



Trinity College Dublin

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

School of Biochemistry and Immunology

**Application of rational computer-aided drug
design approaches to discover novel small
molecule inhibitors of conserved viral non-
structural proteins of SARS-CoV-2**

A PhD thesis submitted to Trinity College Dublin for the
degree of Doctor of Philosophy

✱

2025

Presented by Shubhangi Kandwal (B.pharm, M.Sc.)

Supervisors: Prof. Andrew Bowie and Dr. Darren Fayne

Declaration of authorship

I hereby declare that this thesis represents the sole work of the author except where stated and has not been submitted in whole or in part to any other university as part of a degree or qualification.

The author provides permission for the University's library to lend or copy this work upon request.

The experimental data included in this dissertation were generated by the CACHE organisation and have been used with proper permission and acknowledgment. All analysis, interpretation, and computational/modelling work based on these data were conducted independently by me. All other results and data presented in this dissertation are entirely my own work.

I have read and understood the College / University statement on plagiarism. I declare that the attached dissertation is my own, original work undertaken in partial fulfilment of my degree. I have made no use of sources, materials or assistance other than those which have been openly and fully acknowledged in the text. If any part of another person's work has been quoted, this either appears in inverted commas or (if beyond a few lines) is indented. Any direct quotation or source of ideas has been identified in the text by author, date and page number(s) immediately after such an item, and full details are provided in a reference list at the end of the text. I understand that any breach of the fair practice regulations may result in a failure of this dissertation and that it could also involve other repercussions. I understand also that too great a reliance on the work of others may lead to a low mark.

Student Number: 21203865

Date: 26th August 2025

A handwritten signature in black ink, appearing to read 'Shubhangi', with a long, sweeping horizontal stroke extending to the right.

Shubhangi Kandwal

*In the words of **G.H. Hardy** from the movie **'The Man Who Knew Infinity'**,*

"We are merely explorers of infinity in the pursuit of absolute perfection."

This journey has shown me just how vast and humbling that exploration can be and I'm grateful to all who walked it with me.

Acknowledgements

Firstly, I would like to extend my heartfelt thank you to the Irish Research Council (now known as Research Ireland), which funded the research in this PhD thesis under the grant number **GOIPG/2021/954**.

I am truly grateful to my supervisors, Professor Andrew Bowie and Dr. Darren Fayne, for their constant support and guidance. Their expertise, insightful feedback, and motivation have been truly instrumental in shaping this work. A tremendous thanks to Professor Bowie, it has been a privilege to work under your supervision and to previously have you as a member of my thesis committee.

A special thanks goes to Dr. Darren for believing in me from the very beginning and helping me get through all the challenges that came along the way. Your mentorship has had a huge impact on my growth as a researcher, and I'm truly grateful for everything you've taught me. I cannot forget the annual MDG Christmas dinner, it has been a highlight every year and a great reminder of the supportive and fun community you've built around the MDG group. Thank you for making me understand what science truly is and for making the work enjoyable. It's been an honour to have you as my supervisor and to be part of the MDG group.

I would also like to thank the thesis committee members, Professor Luke O'Neill and Dr. Shigeki Nakagome, for their valuable inputs and discussion sessions during the yearly report presentations. I also thank Drs. Karsten Hokamp and Fiona Roche for their bioinformatics insight and assistance. To the members of the Molecular Design Group (MDG), thank you for all the technical help and for making the research environment enjoyable and motivating. Thank you also to Dr. Ivan (a.k.a. The Python Guru) for patiently dealing with all my silly questions; your help has meant a lot along the way.

I would like to thank the software vendors for their continued support of academic research efforts, in particular the contributions of Chemical Computing Group (CCG), Biovia, and OpenEye Scientific. The support and provisions of Dell Ireland, the Trinity Centre for High Performance Computing (TCHPC), and the Irish Centre for High-End Computing (ICHEC) are also gratefully acknowledged.

A special shoutout and thank you to my close family members Mrs. Renu, Dr. RBS, and Shasha for their endless love, encouragement, and patience throughout my PhD journey. Your belief in me is what kept me going, especially during the toughest moments. You've been my foundation through it all.

I'm also grateful to Mrs. Gayatri (my maasi) for always being there for me with support, advice, and care when I needed it most. Thank you to my late Nani (Mrs. Bhagwati); your love and blessings are with me always. Your wisdom continues to guide me every day. I miss you deeply. This achievement is as much yours as it is mine.

Thanks to my baby sister Nikki for being annoying and somehow always right with your suggestions. Thank you for never getting too mad when I stole your clothes, I owe you more than just a few outfits! I'm so deeply grateful for your love, your honesty, and the way you showed up for me every single time. I couldn't have done this without you.

To that one special person, my partner Shivu, thank you for being my forever constant through it all. I deeply admire how great of a human being you are and how you have stood by me with your unwavering support and care. Your love and belief in me have carried me further than you could imagine. Thank you for all the thoughtful scientific discussions that not only challenged me but also inspired me throughout my PhD. This journey could not have been possible without you, and I dedicate this thesis to you.

Last but not least, a very special thank you to my mom (Mrs. Neelam) and my dad (Mr. Chhatterveer), I owe you everything. Your unconditional love, constant encouragement, and quiet sacrifices have made all of this possible. Thank you for always believing in me when I doubted myself. Whatever that I've been able to achieve is built on the values you raised me with. You guys are the reason I could dream big and reach this milestone. I am, and always will be, proud to be your child. This thesis is dedicated to you.

Abstract

The majority of SARS-CoV-2 therapeutic development work has focussed on targeting the spike protein, viral polymerase and proteases. As the pandemic progressed, many studies reported that these proteins are prone to high levels of mutation and can become drug resistant. Thus, it is necessary to not only target other viral proteins such as the non-structural proteins (NSPs) but to also target the most conserved residues of these proteins. In order to understand the level of conservation among these viruses, we have focussed on the conservation across the coronaviruses and then narrowed our focus to conservation of NSPs across coronaviruses.

A synergistic combination of bioinformatics, computer-aided drug-design and *in-vitro* studies can feed into better understanding of SARS-CoV-2 (SC-2) and therefore help in the development of small molecule inhibitors against the NSPs. As part of our initial anti-viral work, a pharmacophore study on NSP15 found a hit molecule (INS316) that made interactions with Ser293, Lys344 and Leu345 residues (Kandwal & Fayne, 2022) which are highly conserved across SC-2. We have performed multiple sequence alignment studies on different ~1 million sequences of SC-2 NSP sequences to identify the most conserved residues. These residues were then visualized on 3D protein X-ray structures using MOE software. We found that there were known and novel binding pocket residues that were highly conserved in our datasets. Our results indicate that these highly conserved pockets can be targeted for developing promising SC-2 inhibitors. We have also performed mutational analysis and have found different mutational hotspots across the NSPs. Our group was selected to enter two international challenges organized by CACHE to find inhibitors for the RNA binding tunnel of SC-2 NSP13 and the Mac1 domain of SC-2 NSP3. We have used a tiered screening workflow which included the use of volume/shape information of the binding pockets (fastROCS), use of in-house pharmacophore generation software MoPBS/MOE and performed docking in the binding pocket using FRED to rank hits for subsequent clustering and to identify hits that bind to these conserved pockets. These hits were found to make interactions with highly conserved amino acid residues present within the binding pockets of the proteins.

According to the experimental validation performed by the CACHE organisation, it was reported that two of our NSP3 hits showed activity in the HTRF and SPR assays. For the two hits, structural analogues were identified and retested for activity. A working protocol of

Molecular Dynamics (MD) simulation and Absolute Binding Free Energy (ABFE) workflow was developed for the two NSP3 hits using GROMACS. We have also expanded the project to perform hit-to-lead optimization on the two NSP3 hits with the help of the Notre Dame-DCU Seed Grant in Precision Biomedical Technologies funding. Lastly, sequence and structural conservation analysis was performed on Group-IV viruses, primarily comparing helicases and ADP-binding macro domain proteins with SC-2 NSP13 and NSP3, respectively. We found sequence and structural similarity between the Chikungunya virus and SC-2 for helicase and ADP-binding macro domain proteins. This can thus aid in the development of a pan-antiviral inhibitor.

Abbreviations

Abbreviation used	Full form
+ssDNA	Positive sense, single stranded, Deoxyribonucleic acid
+ssRNA	Positive sense, single stranded, Ribonucleic acid
+ssRNA-RT	Positive sense, single stranded, Ribonucleic acid, Reverse transcriptase
3CLpro	chymotrypsin-like protease
AA	Amino acid
ABFE	Absolute Binding Free Energy
ACE-2	Angiotensin-converting enzyme 2
ADMET	Absorption, distribution, metabolism, excretion and toxicity
ADPr	Adenosine diphosphate- ribose
Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asn	Asparagine or aspartic acid
Asp	Aspartic acid
BLOSUM	Blocked Substitution Matrix
CACHE	Critical Assessment of Computational Hit-finding Experiments
CADD	Computer-aided drug design
CGenFF	CHARMM General FF
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CHIKV	Chikungunya virus
Coul	Coulombic
COVID-19	Coronavirus Disease 2019
CoVs	Coronaviruses
CSG	The Coronaviridae Study Group
Cys	Cysteine
DBVS	Docking based virtual screening
DLS	Dynamic light scattering
DMSO	Dimethyl sulfoxide
DPP4	Dipeptidyl peptidase 4
dsDNA	Double stranded, Deoxyribonucleic acid
dsDNA-RT	Double stranded, Deoxyribonucleic acid, Reverse transcriptase
dsRNA	Double stranded, Ribonucleic acid
EC ₅₀	Half maximal effective concentration
Elec	Electrostatic
EndoU	Endoribonuclease

EUA	Emergency use authorization
FDA	Food and Drug Administration
FEP	Free energy perturbation
FFs	Forcefields
FRED	Fast Rigid Exhaustive Docking
GISAID	Global Initiative on Sharing Avian Influenza Data
Gln	Glutamine
Glu	Glutamic acid
Glx	Glutamine or glutamic acid
Gly	Glycine
GTP	Guanosine Triphosphate
HCoV	Human coronavirus
HCV	Hepatitis C Virus
HEL	Helicase
His	Histidine
HMM	Hidden Markov Model
HTRF	Homogeneous Time Resolved Fluorescence
HTS	High-throughput screening
IC ₅₀	Half maximal inhibitory concentration
ICTV	International Committee on Taxonomy of Viruses
Ile	Isoleucine
IUPAC	International Union of Pure and Applied Chemistry
K _d	Dissociation constant
K _i	Inhibition constant
LBDD	Ligand-based drug design
Leu	Leucine
LINCS	Linear Constraint Solver
Lys	Lysine
MD	Molecular Dynamics
MERS-CoV	Middle East respiratory syndrome- coronavirus
Met	Methionine
MHV	Mouse hepatitis coronavirus
MM	Molecular mechanics
MM-GBSA	Molecular Mechanics–Generalised Born/Surface Area
MM-PBSA	Molecular Mechanics Poisson–Boltzmann/Surface Area
MoPBS	Mapping of Protein Binding Sites
Mpro	Main protease
MSA	Multiple sequence alignment

NCBI	National Center for Biotechnology Information.
NiRAN	Nidovirus RdRp-associated nucleotidyltransferase
NJ	Neighbour-Joining
NSP	Non-structural protein
ORF	Open reading frame
PAINS	Pan-assay interference compounds
PAINS	Pan-assay interference compounds
PAM	Point Accepted Mutation
PARP14	Poly(ADP-ribose) polymerase family member 14
PDB	Protein Data Bank
pH	Potential of Hydrogen
Phe	Phenylalanine
PLpro	Papain-like protease
Pro	Proline
QM	Quantum mechanics
QSAR	Quantitative structure-activity relation
RdRp	RNA-dependant RNA-polymerase
RMSD	Root-mean-square deviation
RMSE	Root-mean-square-error
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
ROCS	Rapid Overlay of Chemical Structures
SARS-CoV	Severe acute respiratory syndrome-related coronavirus
SARS-CoV-2	Severe acute respiratory syndrome-related coronavirus 2
SBDD	Structure-based drug design
SC-1	Severe acute respiratory syndrome-related coronavirus 1
SC-2	Severe acute respiratory syndrome-related coronavirus 2
Seq	Sequence
Ser	Serine
SGC	Structural Genomics Consortium
SIMD	single instruction, multiple data
SPR	Surface plasmon resonance
-ssRNA	Negative sense, single stranded, Ribonucleic acid
Thr	Threonine
TI	Thermodynamic integration
TIP3P	Transferable Intermolecular Potential 3 Points
Trp	Tryptophan
Tyr	Tyrosine
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

Val	Valine
vdw	van der Waals
VLP	Virus-like particle
VOC	Variants of concern
VS	Virtual screening
WHO	World Health Organization
WSP	Weighted sum-of-pairs
ZBD	Zinc- binding domain

Table of Contents

Title page	i
Declaration of authorship.....	ii
Acknowledgements	iv
Abstract	vi
Abbreviations.....	viii
Chapter I. Introduction.....	1
1.1. Viruses and their Taxonomic classifications.....	2
1.2. Coronaviridae	4
1.3. Genomic and structural organisation of Coronaviridae	5
1.4. SARS-CoV-2 genomic organisation	6
1.5. SARS-CoV-2 entry into the host cell	13
1.6. Current treatment of SARS-CoV-2 infection.....	14
1.6.1. Protease inhibitors	15
1.6.2. RdRp inhibitors	18
1.6.3. Other inhibitors	21
1.7. Togaviridae	21
1.8. Genomic and structural organisation of Togaviridae	22
1.9. Genomic organisation of Chikungunya virus.....	23
1.10. Replication cycle of Chikungunya virus	25
1.11. Treatment of Chikungunya virus	26
1.12. Bioinformatics-MSA.....	29
1.13. MSA Algorithms	31
1.14. Computer-Aided Drug Design techniques.....	37
1.15. Hypothesis and Objectives	49
1.16. Opportunities	50
Chapter II. Bioinformatics.....	51
2.1. Introduction	52
2.2. Kalign	52
2.3. Conserved regions	54
2.4. Similarities of constituent proteins within RNA viruses	54
2.5. Sequence similarity/identity and protein conservation across coronaviruses.....	56
2.6. Variants of SARS-CoV-2	57
2.7. NSP conservation across coronaviruses	59

2.8. Materials and Methods	67
2.8.1 Bioinformatics materials	67
2.8.2. Benchmarking of MSA tools methods.....	67
2.8.3. NCBI and GISAID databases for downloading sequences	68
2.8.4. Statistical analysis on NCBI dataset	68
2.8.5. Phylogenetic tree on NCBI dataset	69
2.8.6. Pre-processing of NCBI dataset.....	70
2.8.7. Pre-processing of GISAID dataset and statistical analysis.....	70
2.8.8. Variant distribution	72
2.8.9. MSA using Kalign and visualisation of conserved amino acids	72
2.9. Results	74
2.9.1. MSA tools benchmarking results	74
2.9.2. NCBI statistical analysis results	75
2.9.3. Phylogenetic tree of NCBI dataset	76
2.9.4. GISAID statistical analysis.....	78
2.9.5. Variant distribution in NCBI dataset	78
2.9.6. NSP conservation/mutational analysis	80
2.9.7. NSP1	81
2.9.8. NSP2	84
2.9.9. NSP3 Mac1 domain	86
2.9.10. NSP4	88
2.9.11. NSP5	90
2.9.12. NSP6	91
2.9.13. NSP7	93
2.9.14. NSP8	95
2.9.15. NSP9	96
2.9.16. NSP10	98
2.9.17. NSP12	99
2.9.18. NSP13	102
2.9.19. NSP14	104
2.9.20. NSP15	106
2.9.21. NSP16	109
2.10. Discussion	111
2.11. Conclusions and future directions	115

Chapter III. CACHE challenges	116
3.1. Introduction	117
3.2. The CACHE challenges	117
3.3. Conservation analysis used in CACHE challenges.....	117
3.4. Binding sites used for hit finding	118
3.5. CADD techniques	122
3.5.1. Volume based virtual screening (FastROCS).....	122
3.5.2. Pharmacophore based virtual screening.....	124
3.5.3. FRED molecular docking based virtual screening.....	127
3.6. Materials and Methods.....	128
3.6.1. Bioinformatics materials.....	128
3.6.2. Cheminformatics materials	129
3.6.3. Methods for the NSP13 Workflow	129
3.6.4. Methods for the NSP3 Workflow	135
3.7. Results	144
3.7.1. NSP13 results.....	144
3.7.2. NSP3 results.....	176
3.8. Discussion	220
3.9. Conclusions and future directions	225
Chapter IV. Molecular Dynamics simulations & Absolute Binding Free Energy calculations.....	228
4.1. Introduction	229
4.2. Why MD is necessary?.....	229
4.3. MD simulations.....	230
4.4. CHARMM forcefield.....	231
4.5. GROMACS	232
4.6. Absolute Binding Free Energy calculations	234
4.7. Materials and Methods.....	240
4.7.1. Materials.....	240
4.7.2. MD simulations methods	241
4.7.3. ABFE calculations methods.....	244
4.8. Results	247
4.8.1. Benchmarking of MD runs.....	247
4.8.2. Positive control MD runs on SC-2 Mac1-NSP3	248
4.8.3. MD on CACHE SC-2 Mac1-NSP3 hits.....	250

4.8.4. Benchmarking results of ABFE protocol on SC-2 Mpro (NSP5).....	263
4.8.5. ABFE results on SC-2 Mac1-NSP3.....	265
4.9. Discussion	271
4.10. Conclusions and future directions	278
Chapter V. Conclusion and Perspectives	280
Appendix.....	286
6.1. SC-2 Classification chart	286
6.2. Bash commands for SC-2 variant analysis.....	287
6.3. Substructure search of CACHE hits	289
6.3.1. Substructure search of CACHE3I_1689_15	289
6.3.2. Substructure search of CACHE3I_1689_35	290
6.4. GROMACS MD simulation MDP files.....	292
6.5. GROMACS ABFE window spacing.....	296
6.6. Second and third ABFE runs on 5SPT	297
6.7. Second and third ABFE runs on 5SOJ	299
6.8. Second and third ABFE runs on CACHE_1689_35	301
6.9 Publications, presentations, outreach activities, recognition, courses	303
References.....	307

List of Figures

Chapter I. Introduction

1.1. Baltimore classification of viruses, Viral classification of RNA viruses	3
1.2. Classification of Nidovirales.....	4
1.3. Group 4 enveloped and non-enveloped viruses	5
1.4. Cumulative confirmed SC-2 infection cases per million people	6
1.5. Genome of SC-2, structures of NSPs	7
1.6. SARS-CoV-2 entry into the host cell	14
1.7. Classification of Togaviridae family	22
1.8. Genomic organisation of viruses belonging Togaviridae family.....	22
1.9. Global distribution of the CHIKV infection	23
1.10. Genome organisation of CHIKV	24
1.11. Viral replication cycle of CHIKV	26
1.12. Example of a multiple sequence alignment	30
1.13. Workflow behind progressive alignment	32
1.14. Workflow behind iterative alignment (MUSCLE)	34
1.15. Clustal Omega HMM algorithm.....	36
1.16. Classification of CADD	38
1.17. Examples of volume queries of ligand and dummy atoms	40
1.18. Examples of pharmacophore features	42
1.19. Principle of molecular docking	43
1.20. Interactions between protein-ligand complexes	43
1.21. Hydrogen bond.....	44
1.22. MD simulation box	45
1.23. Binding affinity between protein-ligand complex.....	47

Chapter II. Bioinformatics

2.1. Kalign algorithm flow chart	53
2.2. Computational performance of Kalign	54
2.3. MSA of SC-2 variants	58
2.4. NSP2 bound to human 40s ribosomal subunit.....	59
2.5. NSP3 Mac1 MSA of betacoronaviruses	60
2.6. NSP5 MSA of betacoronaviruses and conserved residues at the dimer interface.....	61
2.7. NSP12 MSA of betacoronaviruses	63
2.8. NSP13 MSA of betacoronaviruses	64

2.9. NSP14 MSA of betacoronaviruses.....	65
2.10. NSP15 conserved amino acids in the active site.....	66
2.11. NSP16 conserved amino acids in the active site.....	66
2.12. Benchmarking of MSA tools.....	74
2.13. Benchmarking of MSA tools on 200 sequences.....	75
2.14. Statistical analysis graphs on NCBI dataset.....	75
2.15. Phylogenetic tree of SC-2 variants.....	77
2.16. GISAID sequence submission by different countries.....	78
2.17. Bar graph of distribution of sequences belonging to different variants in the dataset.....	79
2.18. Conservational and mutational analysis of NSP1.....	82
2.19. Binding mode of montelukast with NSP1.....	84
2.20. Conservational and mutational analysis of NSP2.....	84
2.21. Conservational and mutational analysis of NSP3 Mac1 domain.....	87
2.22. Conservational and mutational analysis of NSP4.....	88
2.23. Conservational and mutational analysis of NSP5.....	90
2.24. Conservational and mutational analysis of NSP6.....	92
2.25. Conservational and mutational analysis of NSP7.....	94
2.26. Conservational and mutational analysis of NSP8.....	95
2.27. Conservational and mutational analysis of NSP9.....	97
2.28. Conservational and mutational analysis of NSP10.....	98
2.29. Conservational and mutational analysis of NSP12.....	99
2.30. Conservational and mutational analysis of NSP13.....	102
2.31. Conservational and mutational analysis of NSP14.....	104
2.32. Conservational and mutational analysis of NSP15.....	107
2.33. Conservational and mutational analysis of NSP16.....	110
2.34. SC-2 ORF8 bioinformatics analysis.....	113
Chapter III. CACHE Challenges	
3.1. Domains of SC-2 NSP13.....	118
3.2. SC-2 NSP13 binding sites.....	119
3.3. Schematics of SC-2 NSP3 Mac1 domain role and the binding sites.....	121
3.4. Overview of ROCS algorithm.....	124
3.5. Pharmacophore features.....	125
3.6. FRED docking algorithm.....	127
3.7. Bar graph of amino acid distribution of NSP13 and 3D visualisation of NCBI dataset.....	144

3.8. Bar graph of top 10 mutations from the NCBI SC-2 NSP13 dataset.....	145
3.9. Bar graphs of amino acid distribution and top 10 mutations of GISAID SC-2 NSP13 dataset...	147
3.10. MSA of betacoronaviruses NSP13 protein	150
3.11. MSA of helicase protein of the group IV viruses	152
3.12. Visualization of conserved and mutated amino acids of CHIKV nsP2	153
3.13. PSA between SC-2 NSP13 and CHIKV nsP2.....	154
3.14. Structural overlay between SC-2 NSP13 and CHIKV nsP2	155
3.15. Sequence and structural comparison between SC-2 NSP13 and human RNA helicase.....	158
3.16. RMSD plots of each amino acid residue position of SC-2 NSP13 overlaid structures.....	160
3.17. The RNA binding site/tunnel with siteA1 and siteA2	160
3.18. Conserved amino acids at siteA1 and siteA2.....	162
3.19. Example of fastROCS VS hits of SC-2 NSP13	165
3.20. Example of pharmacophore VS hits of SC-2 NSP13.....	166
3.21. Example of FRED docking VS hits of SC-2 NSP13.....	167
3.22. MDG tiered VS workflow for SC-2 NSP13.....	168
3.23. CACHE hits that passed SPR binding assay against SC-2 NSP13	170
3.24. Steps involved in experimental validation performed by CACHE for challenge 2.....	172
3.25. Bar graph of amino acid distribution of SC-2 Mac1-NSP3 and 3D visualisation of NCBI dataset	177
3.26. Bar graph of top 10 mutations from the NCBI SC-2 Mac1-NSP3 dataset.....	178
3.27. Binding sites of SC-2 Mac1-NSP3.....	179
3.28. Bar graphs of amino acid distribution and top 10 mutations from the GISAID SC-2 Mac-1 NSP3 dataset.....	182
3.29. MSA of betacoronaviruses Mac1 domain protein.....	186
3.30. MSA of Mac1 domain protein of the group IV viruses	187
3.31. Visualization of conserved and mutated amino acids of CHIKV nsP3	188
3.32. PSA between SC-2 Mac1-NSP3 and CHIKV nsP3	189
3.33. Structural overlay between Mac1 domain of SC-2 NSP3 and CHIKV nsP3.....	190
3.34. Sequence and structural overlay of SC-2 Mac1-NSP3 and human PARP14	191
3.35. Conserved ADP-r binding site of NSP3 Mac1 domain	192
3.36. Structural overlay of 22 SC-2 Mac1-NSP3 proteins with residue RMSD plots	193
3.37. Open tunnel and closed tunnel conformation of SC-2 Mac1-NSP3	194
3.38. Conserved amino acids at ADP-r binding site.....	196
3.39. Placement of colour coded dummy atoms on SC-2 Mac1-NSP3.....	197

3.40. Example of fastROCS VS hits of SC-2 Mac1-NSP3	199
3.41. ROC-AUC curve with a score of 0.88.....	202
3.42. Example of FRED docking VS hit of SC-2 Mac1-NSP3.....	202
3.43. MDG tiered VS workflow for SC-2 Mac1-NSP3	203
3.44. Steps involved in experimental validation performed by CACHE for challenge 3	204
3.45. SPR results of SC-2 Mac1-NSP3 hits	208
3.46. ROC-AUC curve of actives and inactives of Mac1 NSP3 from the CACHE experimental validation results.....	214
3.47. Co-relation plots between the experimental logS values and MOE and AqSolPred predicted logS values.....	215
3.48. List of 19 analogues and the parent molecule CACHE3HI_1689_35	217
3.49. Consensus docked pose of RSV-NS1 and STAT1	223
3.50. Structure and docked pose of CACHE_1432_1	224
Chapter IV. Molecular Dynamics simulations & Absolute Binding Free Energy calculation	
4.1. Example of protein-ligand docked pose alongside its MD setup in solution.....	229
4.2. Steps involved in generating MD trajectory	230
4.3. Example of equation used to approximate the atomic forces.....	231
4.4. Flow chart of different stages involved in MD simulation	233
4.5. A typical MD workflow using GROMACS.....	234
4.6. Non-physical thermodynamic cycle	237
4.7. Set of Boresch restraints used for ABFE calculations.....	238
4.8. ABFE alchemical pathway and states controlled by the λ parameters.....	239
4.9. Example of plot showing change in energy during the decoupling of ligand	239
4.10. GROMACS MD simulation workflow.....	244
4.11. Bound and unbound ABFE workflows.....	246
4.12. Benchmarking study plot of MDG servers	247
4.13. Structure of 5SPT with interaction map of S6U and MD RMSD plot	248
4.14. Structure of 5SOJ with interaction map of RWQ and MD RMSD plot	249
4.15. MD analysis results of protein-SK1 complex.....	251
4.16. FRED docked pose and ligand interaction map of SK1	253
4.17. RMSD plot and MD simulation box of protein-SK1 complex	254
4.18. Hydrogen bond analysis of protein-SK1.....	255
4.19. Structural overlay between the MD representative frame and the docked pose of protein-SK1 complex.....	256
4.20. MD analysis results of protein-SK2 complex.....	256

4.21. FRED docked pose and ligand interaction map of SK2	259
4.22. RMSD plot and MD simulation box of protein-SK2 complex.....	260
4.23. Hydrogen bond analysis of protein-SK2	262
4.24. Structural overlay between the MD representative frame and the docked pose of protein-SK2 complex	262
4.25. ABFE results of SC-2 M _{pro} in complex with Z219104216	263
4.26. ABFE results of SC-2 Mac1-NSP3 5SPT	265
4.27. ABFE results of SC-2 Mac1-NSP3 5SOJ	267
4.28. ABFE results of SC-2 Mac1-NSP3 hit CACHE_1689_35	269
4.29. Three docked poses of SK1.....	274
4.30. Predicted binding modes of MERS-CoV NSP5 with IRF3 and KPNA4	275

List of tables

Chapter I. Introduction

1.1. SC-2 NSP's and their functions.....	8
1.2. Protease inhibitors.....	16
1.3. RdRp inhibitors.....	19
1.4. Functions of non-structural proteins of CHIKV	24
1.5. Inhibitors of CHIKV	28
1.6. Different types of sequence alignment algorithms	31
1.7. Different types of MSA algorithms and their corresponding tools.....	31
1.8. Pharmacophore features	41
1.9. List of forcefields used in MD simulations	45

Chapter II. Bioinformatics

2.1. Distribution of sequences belonging to different variants in the dataset.....	79
2.2. NSP sequence lengths along with the type of structure submission.....	80

Chapter III. CACHE Challenges

3.1. Different fragment types used for MoPBS binding site flooding.....	126
3.2. List of conserved amino acids present in the RNA binding site of NCBI dataset	146
3.3. List of conserved amino acids present in the RNA binding site of GISAID dataset.....	148
3.4. Conserved amino acids present in the RNA binding site of betacoronaviruses	150
3.5. List of similar and identical amino acids present in the RNA binding site of SC-2 and CHIKV helicase.....	153
3.6. SiteFinder residues at siteA1 and siteA2.....	161
3.7. Placement of dummy atoms at each pocket of siteA1 and siteA2	163
3.8. List of CACHE hits that passed SPR binding assay against SC-2 NSP13	169
3.9. List of hits with dose response SPR binding assay against SC-2 NSP13	171
3.10. List of hits with aggregation and solubility results for SC-2 NSP13	172
3.11. Docked poses of five hits from round 1 experimental validation.....	173
3.12. List of conserved amino acids present in the Mac1 binding site of NCBI dataset	179
3.13. List of conserved amino acids present in the Mac1 binding site of GISAID dataset.....	182
3.14. Conserved amino acids present in the binding site of NSP3 Mac1 domain of betacoronaviruses	186
3.15. List of similar and identical amino acids present in the binding site of SC-2 and CHIKV Mac1 domain	188
3.16. SiteFinder residues at SC-2 NSP3 binding sites.....	195
3.17. Placement of dummy atoms at each pocket of SC-2 Mac1-NSP3	197

3.18. Examples of pharmacophore hits of each pocket of SC-2 Mac1-NSP3	200
3.19. Compounds that showed >30% inhibition in the primary HTRF assay	206
3.20. DLS solubility and aggregation results of the compounds	206
3.21. SPR dose-response results of the compounds with DMSO concentrations.....	207
3.22. Docked poses of two hits from round 1 experimental validation.....	212
Chapter IV. Molecular Dynamics simulations & Absolute Binding Free Energy calculation	
4.1. Summary of ABFE calculations on the three systems	271
4.2. Predicted binding affinities with gmx_MMPBSA.....	276

Chapter I.

Introduction

Parts of this chapter have been published in:

Kandwal S, Fayne D. Genetic conservation across SARS-CoV-2 non-structural proteins - Insights into possible targets for treatment of future viral outbreaks. *Virology*. 2023 Apr;581:97-115. doi: 10.1016/j.virol.2023.02.011. Epub **2023** Mar 10. PMID: 36940641; PMCID: PMC9999249.

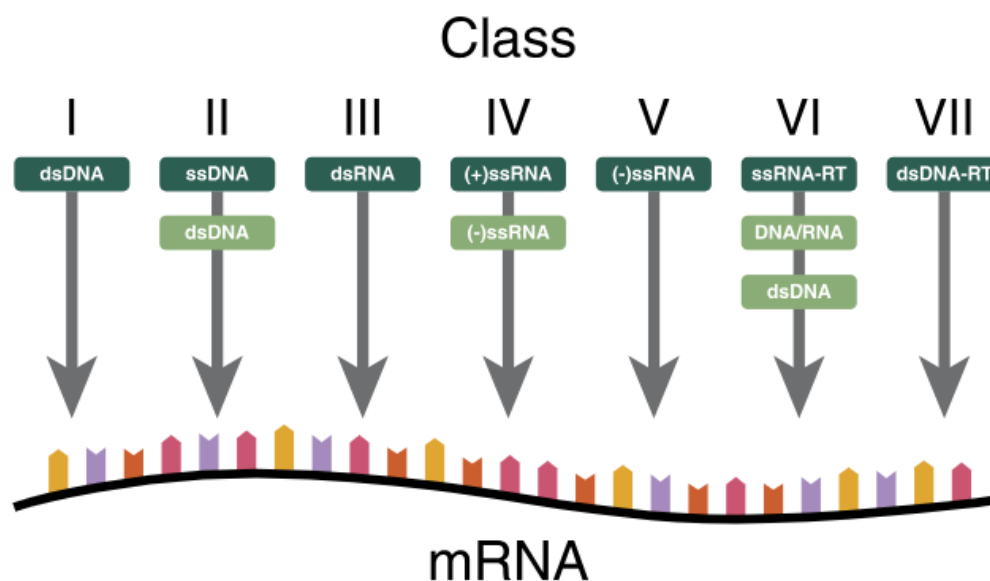
1.1. Viruses and their Taxonomic classifications

Viruses are mobile microscopic parasites containing RNA or DNA genomes which are surrounded by a protective protein coat called the capsid. The nucleic acid material, together with the capsid, is known as a nucleocapsid. For survival, viruses rely on specialised host cells to supply the complex metabolic and biosynthetic machinery of either prokaryotic or eukaryotic cells. The complete infective form of a virus is known as a 'virion' (Gelderblom, 1996). The virion transfers its genetic material into the host cell for transcription and translation by the host cell. For a broader classification, the viruses can be classified into enveloped and non-enveloped viruses. The enveloped viruses have a lipid bilayer derived from the host cell membrane which surrounds the nucleocapsid of the virus, whereas the non-enveloped viruses are known as naked viruses due to the absence of a lipid bilayer. The nomenclature of the virus depends on the type of genomic material i.e., in the case of single stranded RNA, the viruses can be classified into positive sense and negative sense (Gelderblom, 1996).

Viral infections have caused major pandemics in the past, such as the Human Immunodeficiency Virus (HIV) (1981), the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) (SC-1) (2002), the Influenza A virus subtype H1N1 (A/H1N1) (2009), the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) (2012), the Ebola virus (2013), and the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) (SC-2) (2019-present), causing coronavirus disease (COVID-19).

Baltimore (1971) classified viruses on the basis of the type of mRNA produced by the viruses and divided them into seven groups: double-stranded DNA viruses (dsDNA, Group I); single-stranded DNA viruses (+ssDNA, Group II); double-stranded RNA viruses (dsRNA, Group III); positive single-stranded RNA viruses (+ssRNA, Group IV); negative single-stranded RNA viruses (ssRNA, Group V); positive single-stranded RNA viruses with DNA intermediates (+ssRNA-RT, Group VI), commonly known as retroviruses; and the double-stranded DNA retroviruses (dsDNA-RT, Group VII) (Figure 1.1. a).

(a)



(b)

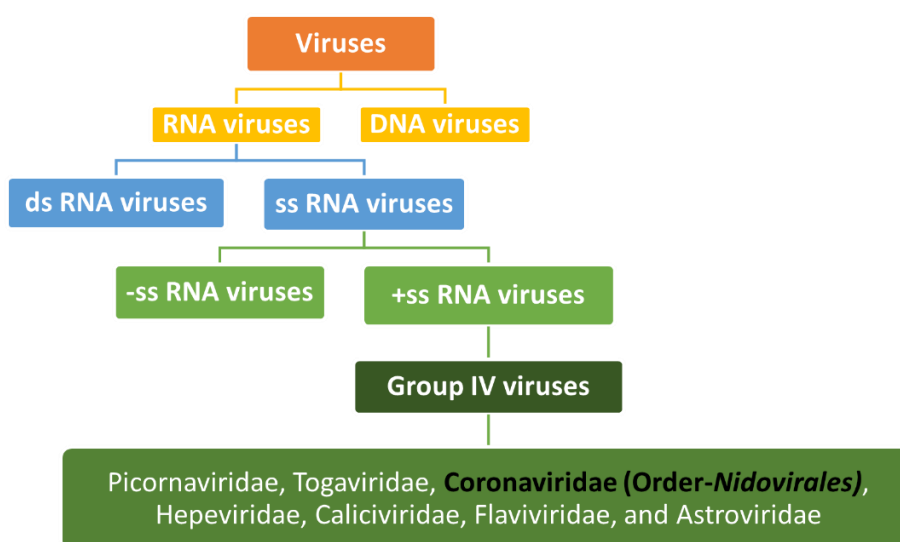


Figure 1.1. (a) Baltimore classification of viruses based on the genetic material and how they produce the mRNA. Group IV viruses consist of (+) sense single stranded (ss) RNA and later also produces an intermediate (-) sense single stranded RNA. (b) Viral classification of RNA viruses focusing on the Group IV viruses.

As per the Baltimore classification of viruses, the Group IV viruses have a positive-sense RNA which can be translated directly into protein as it serves as a template for RNA-dependent RNA polymerase, yielding a complementary RNA strand to form a dimer with the template strand. This double-stranded RNA then serves as the template for a new positive-sense genome which is independently infectious. The Group IV viruses comprise

seven sub-classes of viruses that are either enveloped (*Flaviviridae*, *Togaviridae* and *Coronaviridae*) or non-enveloped (*Caliciviridae*, *Picornaviridae*, *Hepeviridae* and *Astroviridae*) (Berman, 2012) (Figure 1.1. b).

1.2. Coronaviridae

The order *Nidovirales* consists of RNA viruses having linear genome sizes of 26–32 kb, includes coronaviruses, toroviruses and roniviruses (Gorbalenya et al., 2006). The coronaviruses (CoVs) are enveloped positive-strand RNA viruses that can cause life-threatening disease in humans. There are 39 species in 27 subgenera, five genera and two subfamilies that belong to the family *Coronaviridae*. The coronaviruses are spherical in shape having a diameter of 80–160 nm. The *Coronaviridae* family is divided into subfamilies *Torovirinae* and *Coronavirinae*. Genetic and antigenic characteristics are used to divide coronaviruses into four genera: α , β , γ , and δ (Figure 1.2.). Further classification is shown in Appendix 6.1.

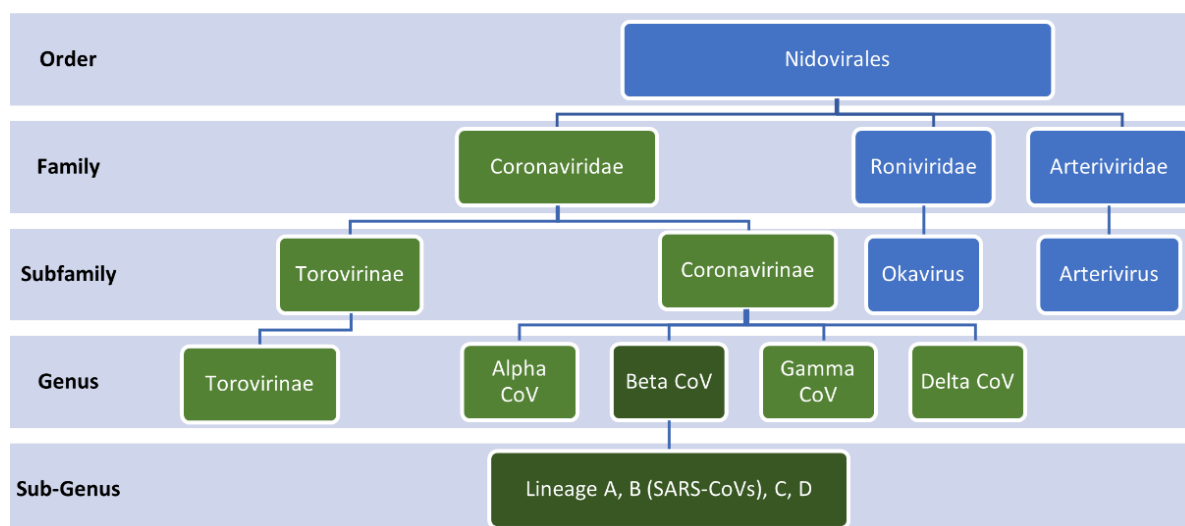


Figure 1.2. Classification of sub-genus, genus, subfamily and family that falls under the order *Nidovirales*.

Mammals are mostly infected by α and β coronaviruses whereas γ and δ tend to infect birds and some can also infect mammals (Woo et al., 2012) (Cui et al., 2019). Coronaviruses are known for causing common colds and have gathered a lot of attention due to their ability to cause zoonotic infections (Pyrce et al., 2007) (Graham et al., 2013). The 2003 outbreak of SC-1 and the 2012 outbreak of MERS-CoV were caused by betacoronaviruses found in bats and other species prior to human transmission (Ge et al., 2013; Menachery et al., 2015). Within the subgenus lineage B, SC-1 and SC-2 phylogenetically share a most recent

common ancestor and are relatively distant to MERS-CoV (belonging to the subgenus *Merbecovirus*) in the genus *Betacoronavirus* (Chen et al., 2021).

1.3. Genomic and structural organisation of Coronaviridae

The group 4 viruses have +ssRNA genomes that serve as mRNA. A negative strand is synthesised to generate a double stranded replicative form (dsRF) which is used to make multiple copies of the genome. The viruses within this group are enveloped and non-enveloped (Figure 1.3.).

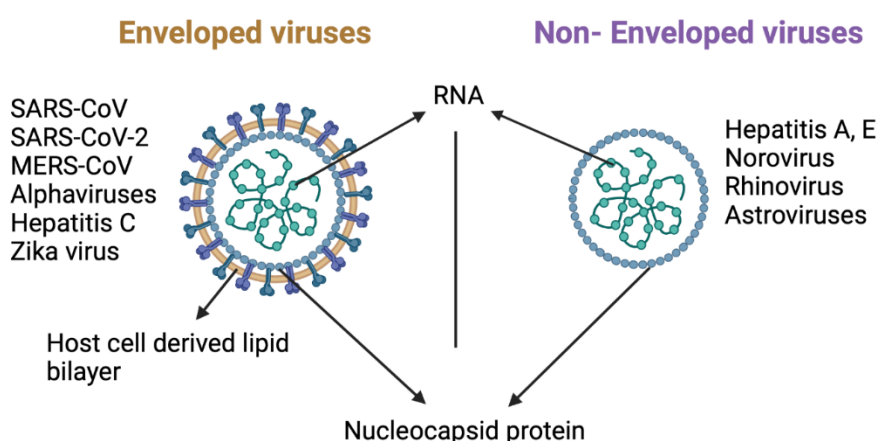


Figure 1.3. Illustration of Group 4 enveloped and non-enveloped viruses. The enveloped viruses comprise host cell derived lipid bilayer (yellow), nucleocapsid protein (blue beads), the spike protein (blue spikes) and RNA. The non-enveloped viruses comprise only the nucleocapsid protein and the RNA. Created with BioRender.

Viruses such as SC-1, mouse hepatitis coronavirus (MHV), MERS-CoV, bovine coronavirus (BCoV), bat coronavirus HKU4 and human coronavirus OC43, including SC-2, belong to β -coronaviruses (Fehr and Perlman, 2015). The infection caused by these coronaviruses are a result of zoonotic transmission and spread among humans through close contact. Many coronaviruses that infect humans, such as HCoV-NL63, HCoV-229E, HCoV-OC43 and HCoV-HKU1, are known to cause mild clinical symptoms, however SC-1, MERS-CoV, and the newly identified SC-2 infection cause severe respiratory illness (Xu et al., 2020). COVID-19 was first identified in December 2019 in Wuhan City, China (Hui et al., 2020), negatively impacting the lives of the global population due to the resulting healthcare, economic, social and mental health crises. There have been approximately 704 million COVID-19 cases around the world as of 17th June 2025 (Worldometer, 2020 COVID - Coronavirus Statistics

Worldometer (worldometers.info)). As of 13th July 2025, the cumulative confirmed SC-2 infection cases per million people is shown in Figure 1.4.

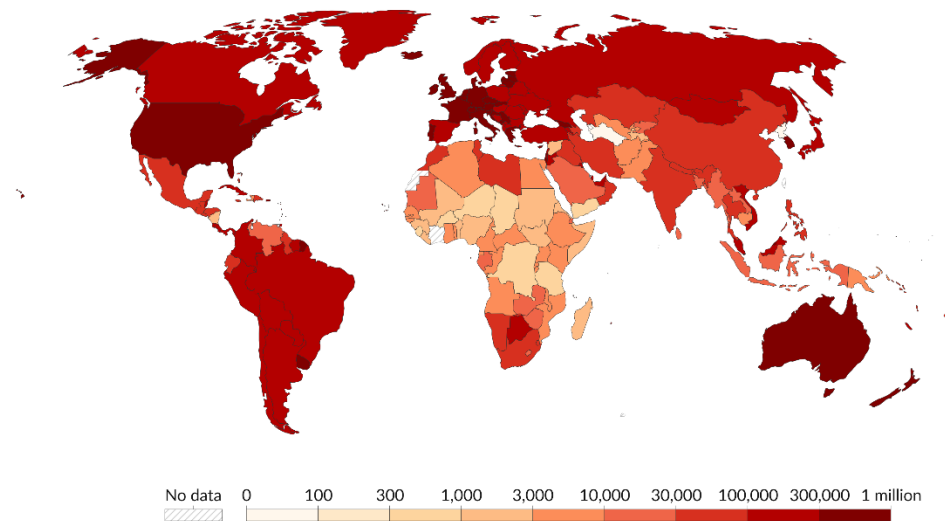


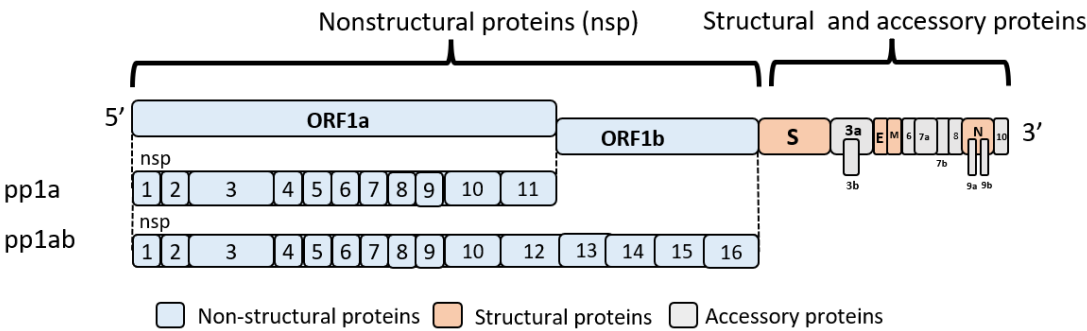
Figure 1.4. Cumulative confirmed SC-2 infection cases per million people. The colour scale represents infected cases per million people. The red colour indicates a high number of infection cases, whereas the cream colour signifies low levels of infection cases. Schematic taken from Coronavirus (COVID-19) Cases - Our World in Data.

1.4. SARS-CoV-2 genomic organisation

The SC-2 genome encodes for structural, accessory and NSPs. There are ~30,000 nucleotides within the SC-2 genome and it possesses 14 ORFs encoding for at least 27 proteins (Wu et al., 2020). The translation of these Open Reading Frames (ORFs) by the host cell machinery give rise to polyproteins that are processed by the host and viral proteases (Modrow et al., 2013). The viral RNA is translated from two open reading frames ORF1a and ORF1ab which thus produces two large polyproteins, PP1a and PP1ab (Kim et al., 2020, Kim et al., 2020; Tok and Tatar, 2017). Cleaving of the polyproteins is achieved by two viral proteases that are present within the polyprotein structure. These are papain-like protease (PL^{pro}, NSP3) and chymotrypsin-like protease (3CL^{pro}, also called M^{pro}, NSP5), which produce the mature NSPs (Ghosh et al., 2020) (Figure 1.5.). There is production of 10 fragments, constituting NSP1-NSP10 when PP1a is cleaved by the protease. On the other hand, PP1ab is cleaved to produce all the NSPs i.e., NSP1-NSP16. A component of NSP3 known as the PL^{pro} protease is involved in the formation of NSP1, NSP2 and NSP3. The remaining NSPs (NSP4-NSP16) are processed by 3CL^{pro} which is present

within NSP5 (Thiel et al., 2003). These NSP's are necessary for the genome replication and some of their functions are mentioned in Table 1.1. along with sequence similarity percentage between SC-1 and SC-2. All these NSP's are common across members of the Beta-CoV genera.

(a)



(b)

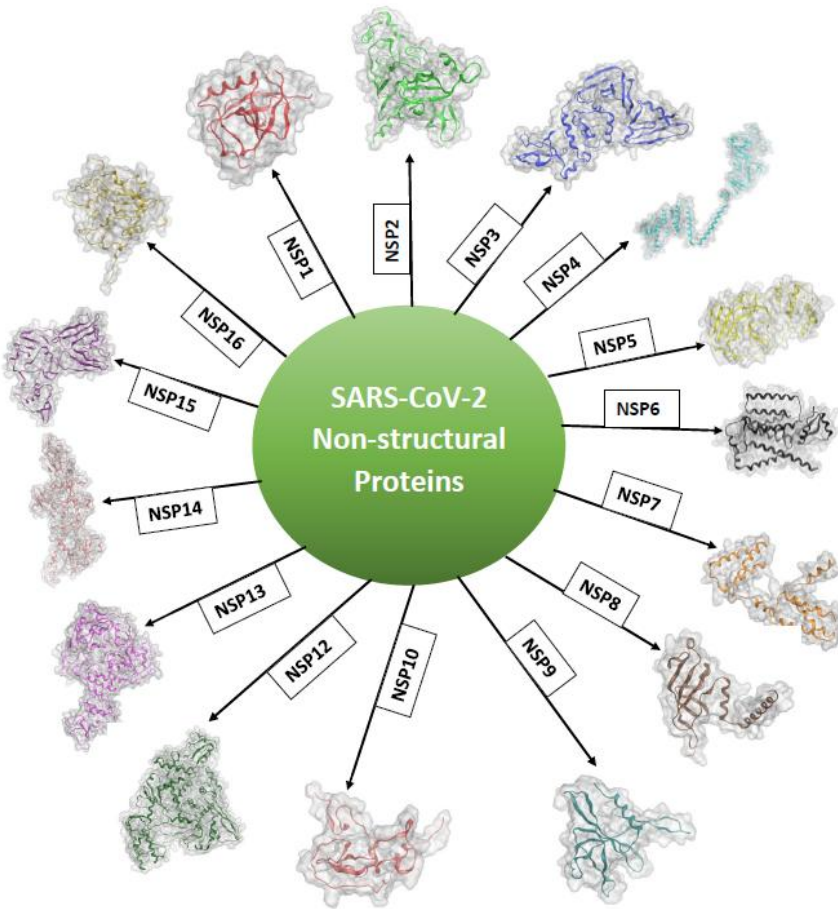


Figure 1.5. (a) The SC-2 genome encodes the long polyprotein-1ab which is cleaved into 16 NSPs (blue) in addition to structural (orange) and accessory proteins (grey). (b) Known structures of SC-2 NSPs.

Amino Acid Length	NSP name, accession number and PDB ID	Description	Sequence similarity between SC-1 and SC-2
180	NSP1 YP_009725297.1 PDB- 7K3N (17 structures)	Leader protein known to suppress host antiviral response. Involved in mRNA degradation (Almeida et al., 2007).	91.1%
638	NSP2 YP_009725298.1 PDB- 7EXM (3 structures)	Interacts with host prohibitins (PHB, PHB2) thus modulating host cell survival signalling pathway (Harcourt et al., 2004).	82.9%

1945	NSP3 YP_009725299.1 PDB- 6W9C (362 structures)	Presence of several domains. Papain-like proteinase (PLpro) functions as protease to cleave the translated polyprotein. NSP3-NSP4-NSP6 complex responsible for viral replication (Neuvonen and Ahola, 2009).	86.5%
500	NSP4 YP_009725300.1	Contains transmembrane domain 2. Anchors the viral replication-transcription complex to the modified ER membrane. NSP3-NSP4-NSP6 complex (Manolaridis et al., 2009).	90.8%
306	NSP5 YP_009725301.1 PDB- 7BQY (505 structures)	3C-Like proteinase/main proteinase (3CLpro/Mpro). Cleaves viral polyprotein during replication. Also blocks host innate	98.7%

		immune response (Anand et al., 2002).	
290	NSP6 YP_009725302.1	Has a transmembrane domain. Initial induction of autophagosomes from host ER. NSP3-NSP4-NSP6 complex (Gordon et al., 2020).	94.8%
63	NSP7 YP_009725303.1 PDB- 7BV2 (48 structures)	Hexadecameric NSP7-NSP8 complex part of RdRp (Peti et al., 2005).	100.0%
198	NSP8 YP_009725304.1 PDB- 7BV2 (52 structures)	Hexadecameric NSP7-NSP8 complex part of RdRp (Peti et al., 2005).	99.0%

113	NSP9 YP_009725305.1 PDB- 6W9Q (11 structures)	Acts as ssRNA-binding protein thus helps in viral replication (Egloff et al., 2004)	98.2%
139	NSP10 YP_009725306.1 PDB- 6W4H (43 structures)	Plays a role in stimulating NSP14 3'-5' exoribonuclease and NSP16 2'-O-methyltransferase activities (Joseph et al., 2006).	99.3%
13	NSP11 YP_009725312.1	First few amino acid of NSP12	76.9%
932	NSP12 YP_009725307.1 PDB- 7BV2 (43 structures)	RdRp activity essential for viral replication and transcription (Zanotto et al., 1996).	98.1%
601	NSP13 YP_009725308.1 PDB- 6XEZ (71 structures)	Helicase/triphosphates; catalyses the unwinding of duplex nucleic strands thus helps in translation (Nishikiori et al., 2012).	100.0%

527	NSP14 YP_009725309.1 PDB- 7EGQ (64 structures)	Proofreading Exoribonuclease domain (3'-5' exonuclease) (ExoN). Prevents errors during viral genome replication (Gordon et al., 2020).	98.7%
346	NSP15 YP_009725310.1 PDB- 6WLC (48 structures)	EndoRNAse; Uridine- specific endoribonuclease (NendoU) (Gordon et al., 2020).	95.7%
298	NSP16 YP_009725311.1 PDB- 6W4H (19 structures)	RNA-cap methyltransferase; 2'-O-ribose methyltransferase; cap methylation of the mRNA; helps in evading the immune system (von Grotthuss et al., 2003)	98.0%

Table 1.1. Known NSP's with their functions, amino acid length, accession number and Protein Data Bank (PDB) IDs (example PDB ID and the total number of PDB structures as on July 2022) along with percentage similarity between SC1 and SC-2. The pairwise sequence alignment was performed using EMBOSS Needle at <https://www.ebi.ac.uk> (Madeira et al., 2022).

The pp1ab polyprotein is encoded by the ribosomal frameshift mechanism of gene 1b. pp1ab has 7096 amino acid residues and is known to be well conserved across CoVs (Naqvi et al., 2020). Apart from the structural proteins and NSPs, the SC-2 genome also encodes for eleven accessory proteins: ORF3a, ORF3b, ORF3c, ORF3d, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF9c and ORF10 (Shang et al., 2021). These accessory proteins play an important role in pathogenesis but further discussion of most of them is outside the scope of this work.

1.5. SARS-CoV-2 entry into the host cell

In the SC-2 genome there are structural proteins such as spike (**S**), envelope (**E**), membrane (**M**) and nucleocapsid (**N**) proteins located at the 3' terminus. These proteins share high sequence similarity with the corresponding protein of SC-1 and MERS-CoV (Naqvi et al., 2020). The spike protein of CoVs facilitates binding to the host cell-surface receptor during host cell entry (Lau and Peiris, 2005). The receptor-binding domain (RBD) in the S1 subunit of the S protein is responsible for binding to the host receptor thus resulting in fusion of the S2 subunit to the host cell membrane. SC-1 and SC-2 recognize the human ACE2 receptor to bind with the viral S protein followed by entry into the host cell as shown in (Figure 1.6.) (Zhang et al., 2020, Zhang et al., 2020). In addition to ACE2, coronaviruses utilise different receptors such as aminopeptidase N (CD13) (Yeager et al., 1992), 9-*O*-Acetylate sialic acid (Huang et al., 2015) and DPP4 (Wang et al., 2013) to facilitate entry into the host cell.

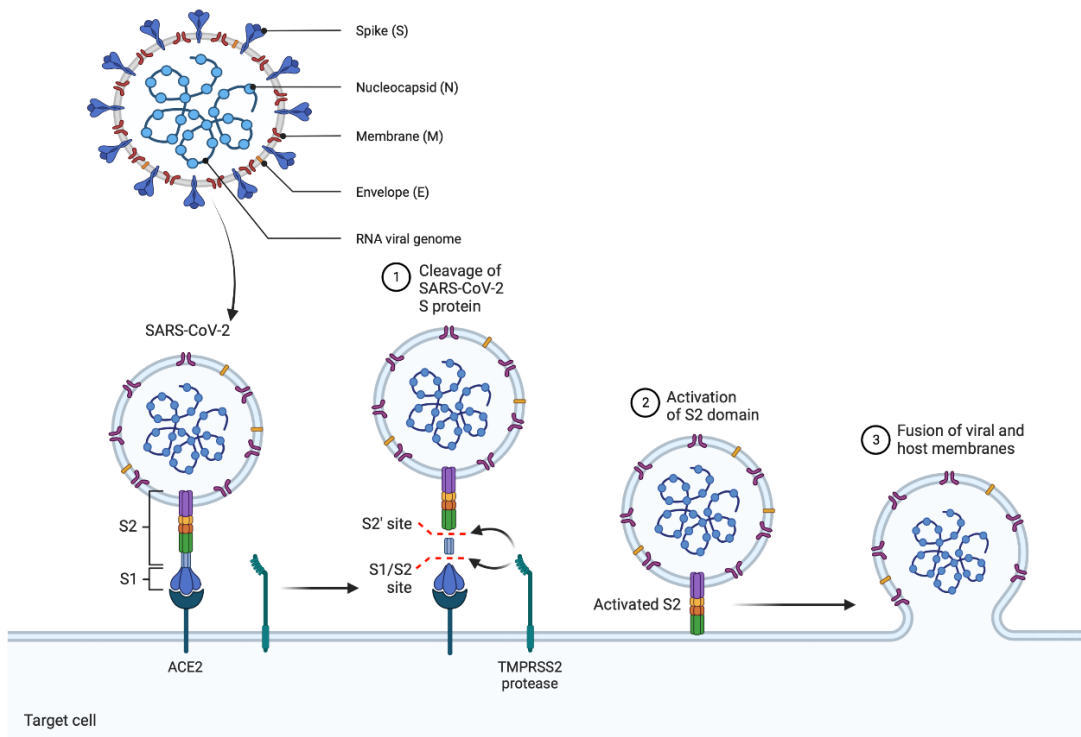


Figure 1.6. SC-2 spike protein interacting with host ACE-2 receptor to facilitate entry into the host cell. Upon the attachment of the SC-2 spike protein to the ACE-2 receptor, the S1/S2 site is cleaved by the TMPRSS2 protease, which results in the activation of the S2 domain. The activation of the S2 domain facilitates the fusion of the virus and the host membrane. Schematic taken from “Mechanism of SC-2 Viral Entry” and “Human Coronavirus Structure”, by BioRender (2022). Retrieved from <https://app.biorender.com/biorender-templates>.

1.6. Current treatment of SARS-CoV-2 infection

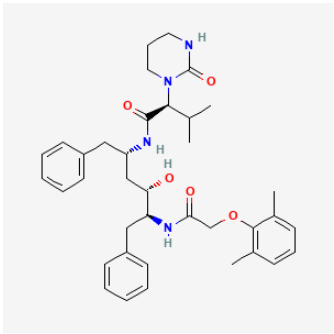
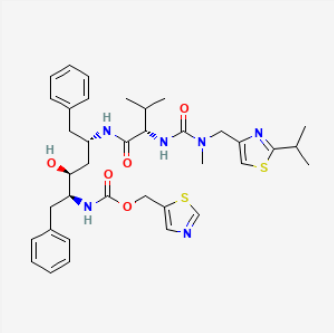
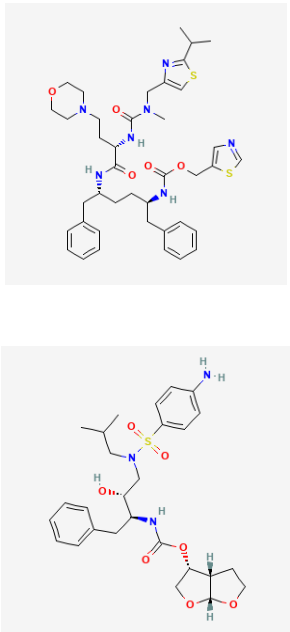
The infection caused by SC-2 in humans can cause mild, moderate or severe symptoms. After the viral spike protein binds with the epithelial cells in the respiratory tract, the virus enters the cells and starts replicating, migrating down to the airways thus entering alveolar epithelial cells in lungs. The immune response is triggered due to rapid replication of viral particles in the lungs. There can be respiratory failure and respiratory distress due to the production of a cytokine storm. This cytokine storm syndrome is considered to be a major cause of death in patients with COVID-19 (Huang et al., 2020). In severe cases of COVID-19, life-threatening pneumonia develops. All ages of the population are susceptible to COVID-19 however patients over the age of 60 years old with co-morbidities are more likely to develop severe respiratory disease (Guan et al., 2020).

There are a variety of therapeutic modalities available under FDA issued Emergency Use Authorization (EUA) or being evaluated for the management of COVID-19; for example, antiviral drugs (e.g., molnupiravir, paxlovid or remdesivir), anti-SC-2 monoclonal antibodies (e.g., bamlanivimab/etesevimab or casirivimab/imdevimab), anti-inflammatory drugs (e.g., dexamethasone) and immunomodulators agents (e.g., baricitinib or tocilizumab) (Coopersmith et al., 2021). To contain the global pandemic, vaccination is an important approach for preventing and slowing down the transmission of SC-2. Researchers all around the world have worked on developing novel vaccines against SC-2 and according to ourworldindata.org 13.72 billion doses of vaccine doses have been administered as of 12th August 2025.

For treatment purposes, there are many research studies which are focusing on designing and developing antivirals specific to SC-2. Conserved enzymes such as 3CLpro/Mpro (NSP5), papain-like protease (PLpro) (NSP3) and RdRp (NSP12) are promising drug targets as these enzymes have sequence similarities with viruses like SC-1 and MERS-CoV (Zumla et al., 2016). It can be hypothesised that due to the sequence similarities between SC-1, MERS-CoV and SC-2 that therapeutic molecules used for targeting SC-1 and MERS-CoV could also prove to be useful for SC-2 with a similar efficacy. The search for antiviral drugs has primarily concentrated on repurposing existing drugs, for example remdesivir, which is an antiviral drug acting on HIV reverse transcriptase and was approved for emergency treatment of COVID-19 patients (Parks and Smith, 2020).

1.6.1. Protease inhibitors

Viral proteases are productive targets for antiviral therapies. The virtual screening of HIV protease inhibitors lopinavir, ritonavir, darunavir and cobicistat (Table 1.2.) against SC-2 3CLpro (NSP5) predicted inhibition of the protease. Lopinavir showed a significantly inhibitory effect on SC-2 in Vero E6 cells with an EC₅₀ of 26.63 μ M (Choy et al., 2020). A group in the USA used a docking approach and found that the anticoagulant agent dipyridamole inhibited Mpro as determined by the surface plasmon resonance assay *in vitro*. More than 50% replication of SC-2 was inhibited at a concentration of 100 nM in Vero E6 cells (Liu et al., 2020).

Drugs	Structure	Status
Lopinavir		Approved for HIV In phase 4 trials of SC-2 (NCT04255017)
Ritonavir		Approved for HIV In phase 3 trials for SC-2 (NCT04261270)
Darunavir and cobicistat		Approved for HIV In phase 3 trials for SC-2 (NCT04252274)

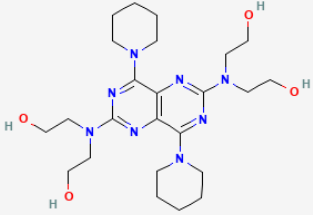
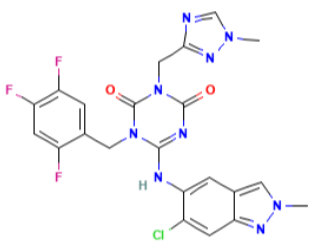
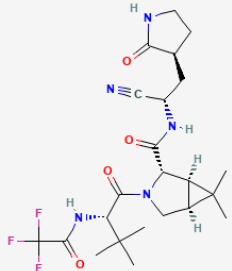
Dipyridamole		In phase 2 trials for SC-2 (NCT04424901)
S-217622 (Emsitrelvir)		In phase 3 trials for SC-2 (jRCT2031210350)
Nirmatrelvir (Paxlovid)		Completed phase 4 trials for SC-2 (NCT05366192)

Table 1.2. The structure and clinical trial status of drugs having inhibitory action against SC-2 3CL^{pro} (NSP5) currently being investigated.

SC-1 PL^{pro} has protease and phosphatase activity and contributes to the viral replication process (Báez-Santos et al., 2015). It is responsible for cleaving NSP1, NSP2 and NSP3 from the large polyprotein (Rut et al., 2020). A study determined that 6-mercaptopurine ($IC_{50} = 21.6 \mu M$) and 6-thioguanine ($IC_{50} = 5 \mu M$), which are used for the treatment of childhood acute lymphoblastic leukaemia and myeloid leukaemia respectively, showed inhibitory activity against SC-1 PL^{pro} (Chou et al., 2008). Among these drugs, 6-thioguanine

inhibited SC-2 viral replication in Vero-E6 cells with a EC_{50} value of $0.64 \pm 0.37 \mu M$ (Swaim et al., 2020).

An orally administered drug known as S-217622 (ensitrelvir) which is also a SC-2 M^{pro} inhibitor decreases viral load and ameliorates disease severity in SC-2 infected hamsters (Sasaki et al., 2023). It also inhibited viral proliferation in Vero E6 cells with an EC_{50} value of $0.35 \mu M$ (Unoh et al., 2022).

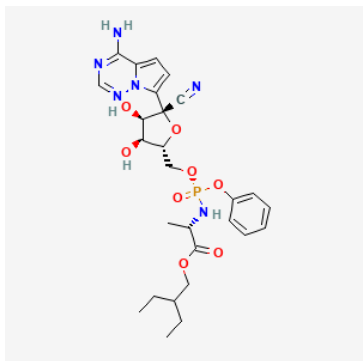
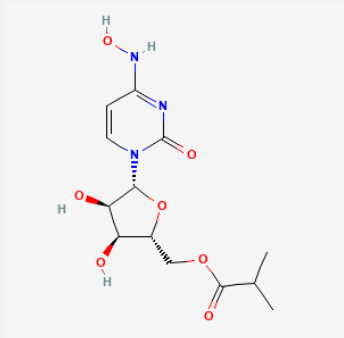
The FDA has authorized Paxlovid (nirmatrelvir and ritonavir) in December 2021 (M^{pro} inhibitor) under emergency use authorization against SC-2 infection which is used for the treatment of mild-to-moderate COVID-19 in adults and paediatric patients (12 years of age and older weighing at least 40 kg). A recent study suggests that Paxlovid is highly effective at reducing the risk of severe COVID-19 or mortality. In the study 4737 patients were treated with Paxlovid who had adequate COVID-19 vaccination status. Both were associated with significant decrease in the rate of severe COVID-19 or mortality (Najjar-Debbiny et al., 2022). Heilmann et al. (2023) identified several mutations such as Y54C, G138S, L167F, Q192R, A194S and F305L in the SARS-CoV-2 M^{pro} that confer resistance to the M^{pro} inhibitors nirmatrelvir and ensitrelvir.

1.6.2. RdRp inhibitors

NSP12 is where the RdRp activity of the virus resides (Koonin et al., 1989). This protein is necessary for the viral replication process, so it is a promising anti-viral drug target. NSP12 is targeted by small molecule inhibitors like remdesivir (GS-5734), favilavir (T-705) and ribavirin (Ul Qamar et al., 2020). Among these remdesivir has shown efficacy against SC-2 *in vitro* and *in vivo* (Williamson et al., 2020). Contrary to this, a large-scale study conducted on 5000 participants by the World Health Organization's (WHO) Solidarity trial consortium found that remdesivir had little or no effect on hospitalized patients with COVID-19 indicated by overall mortality, initiation of ventilation, and duration of hospital stay (WHO Solidarity Trial Consortium, 2021).

Favilavir, which is an antiviral drug used for treating influenza, has been approved in China, Russia and India for the treatment of COVID-19. The time of viral clearance was reported

to be reduced by the use of favilavir in a clinical study conducted in China (Cai et al., 2020). In December 2021 the FDA issued an EUA for Merck's oral pill molnupiravir (NSP12 inhibitor) for treatment of SC-2 infection. A recent study conducted on 40,776 patients infected with SC-2 during the BA.2 wave found that initiation of novel oral antiviral treatments such as molnupiravir or nirmatrelvir–ritonavir in hospitalised patients not requiring oxygen therapy on admission showed substantial clinical benefit (Wong et al., 2022). Taking the example of HCV (group 4 virus), the RdRp (NS5B polymerase) is targeted by sofosbuvir (Table 1.3.) which is a nucleotide analogue inhibitor (Summers et al., 2014). A study by Sacramento et al. (2021) found that sofosbuvir alone and combined with daclatasvir inhibited replication of SC-2 in Calu-3 cells with EC_{50} value ranging from 0.5 to 0.7 μ M.

Drug (alternative names)	Structure	Status
Remdesivir (GS-5734; Veklury)		Approved for marketing (NCT04323761)
Molnupiravir (EIDD-2801)		FDA EUA following phase 3 trials for SC-2 (NCT04939428, NCT05271929)

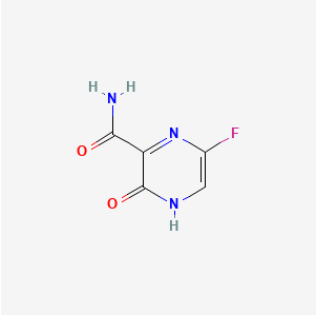
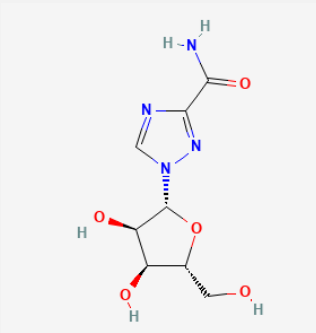
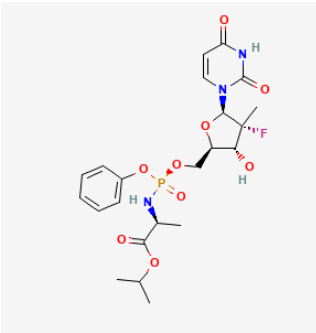
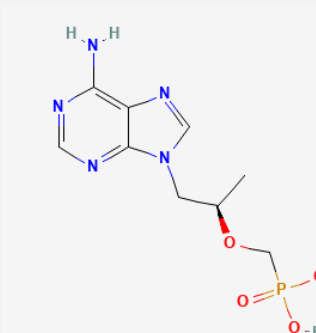
Favipiravir (Avigan; Favipira; T-705)		In phase 4 trials for SC-2 (NCT04359615)
Ribavirin (Virazole)		In phase 3 trials for SC-2 (NCT04392427)
Sofosbuvir (Sovaldi)		Completed phase 4 (NCT04498936, NCT04773756)
Tenofovir (TDF/Viread, TAF/Vemlidy)		In phase 3 trials for SC-2 (NCT04405271, NCT04334928- completed, NCT04890626)

Table 1.3. The structure and clinical trial status of drugs having inhibitory action against SC-2 RdRp (NSP12) currently being investigated.

1.6.3. Other inhibitors

The SC-2 NSP13 is the helicase protein which separates and rearranges the viral nucleic duplex prior to transcription and replication (Kwong et al., 2005). The helicase of SC-2 has 99.8% sequence identity to SC-1 NSP13 helicase (White et al., 2020). Some reported inhibitors include bananins, 5-hydroxychromone derivatives, ADKs and SSYA10-001, which have potential to be used in the treatment of COVID-19 as they had an inhibitory effect on SC-1 (Tanner et al., 2005; Kim et al., 2011; Lee et al., 2009; Adedeji et al., 2012). Among these compounds SSYA10-001 presented weaker antiviral activity against SC-2 in Vero E6 cells with an EC₅₀ value of 81 µM. In the same study another compound FPA-124 was tested for antiviral activity against NSP13 of SC-2 in Vero E6 cells giving an EC₅₀ value of 14 µM (Zeng et al., 2021).

Another interesting protein target is 2'-O-Ribose methyltransferase (2'-O-MTase), known as NSP16 which facilitates the methylation of the 5' terminal of the mRNA thus shielding it from degradation by the host immune system (Decroly et al., 2011; Yoneyama and Fujita, 2010). A study by Bobrovs et al. (2021) found three potent SAM analogues ZINC4228245 (IC₅₀ 1.02 µM), ZINC3861767 (IC₅₀ 1.14 µM) and ZINC473112262 (IC₅₀ 2.46 µM) acting against NSP16 of SC-2 in A549 human non-small cell lung cancer cell line, HepG2 hepatocarcinoma cells and BALB/c 3T3 clone A31 murine fibroblast cell lines.

1.7. Togaviridae

The family *Togaviridae* consists of small, enveloped viruses having single-stranded, positive-sense RNA genomes of about 10-12 kb. The family can be classified into two genus *Alphavirus* and *Rubivirus* (Chen et al., 2018). The genus *Alphavirus* consist of many diverse species whereas the genus *Rubivirus* only includes the single species *Rubella virus* (Figure 1.7.). These viruses are also a part of the Group-IV viruses. The *Alphavirus* genus are arthropod-borne virus species known to infect vertebrates like humans and are widely present around the globe. These viruses are pathogenic in humans causing arthritis, encephalitis, rash, and fever (Strauss & Strauss, 1994) (Kubes & Ríos, 1939). For these viruses small rodents and birds are natural reservoirs and mosquitos serve as vectors (Guerrero-Arguero et al., 2021). The Alphaviruses are sub-classified into Old World (OW)

and New World (NW) groups. The OW group consists of *Sindbis*, *Chikungunya*, *Ross River* and *Semliki Forest virus* whereas NW group consists of *Eastern Equine Encephalitis*, *Western Equine Encephalitis* and *Venezuelan Equine Encephalitis virus*.

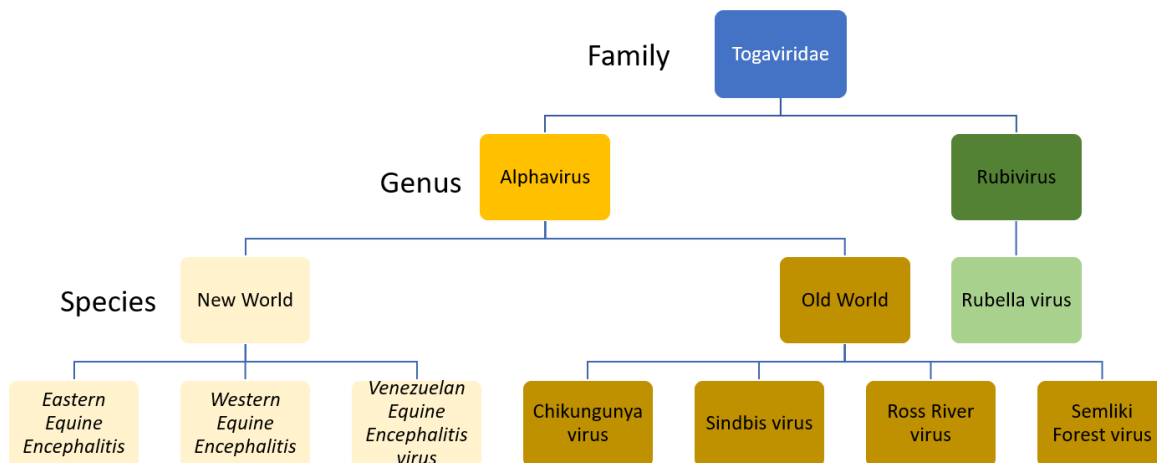


Figure 1.7. The classification of the Togaviridae family based on genus and species.

1.8. Genomic and structural organisation of Togaviridae

The virions of viruses belonging to the family *Togaviridae* are spherical, enveloped and about 70nm in diameter. They consist of a nucleocapsid core which is surrounded by a lipid bilayer embedded with glycoprotein spikes (Cheng et al., 1995) (R. Zhang et al., 2011) (Figure 1.8.). The genome of these viruses is capped and has a poly-A tail.

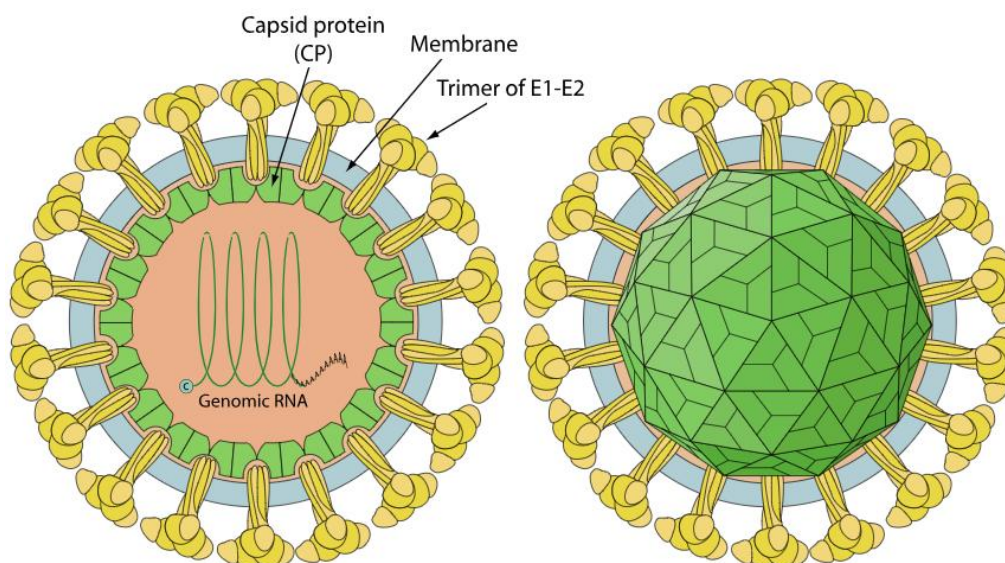


Figure 1.8. Genomic organisation of the viruses belonging to family Togaviridae. The trimer envelope (E1-E2) is represented in yellow, the membrane protein is represented

in blue, and the capsid protein is represented in green. The virus encapsulates the genomic RNA. Schematic taken from ViralZone.expasy.org.

1.9. Genomic organisation of Chikungunya virus

Chikungunya virus (CHIKV) is an *Alphavirus* within the family of *Togaviridae* which is transmitted to humans by the *Aedes* species of Mosquitos (Kendall et al., 2019). The virus is known to be responsible for causing periodic and explosive outbreaks characterised by severe or prolonged polyarthrititis (Silva & Dermody, 2017). The CHIKV has previously been reported to have undergone epidemic spread particularly in the areas of Africa and South-East Asia, India, China, Central and South America and the Caribbean since ~2000 (Fischer & Staples, 2014) (Wahid et al., 2017). The symptoms of the infection include fever, rash, myalgia, polyarthralgia and headache (Silva & Dermody, 2017). The global distribution of the CHIKV infection is represented in Figure 1.9.

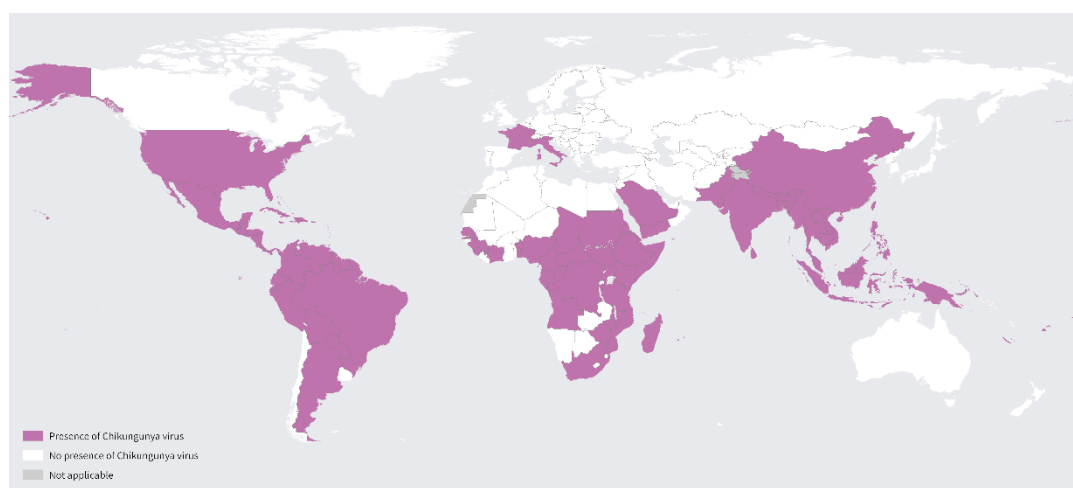


Figure 1.9. The global distribution of CHIKV infection. The presence of the CHIKV virus is indicated in purple, while its absence is shown in white. The geographical region shaded in grey is the region which is not applicable. Schematic taken from the World Health Organisation (WHO).

The genome of the CHIKV is 11,805 nucleotides long and is known to encode two polyproteins: the non-structural polyprotein which consist of four proteins (nsP1, nsP2, nsP3 and nsP4) (Table 1.4.) and the structural polyprotein which consist of five proteins (capsid, E3, E2, 6K and E1) (Figure 1.10.).

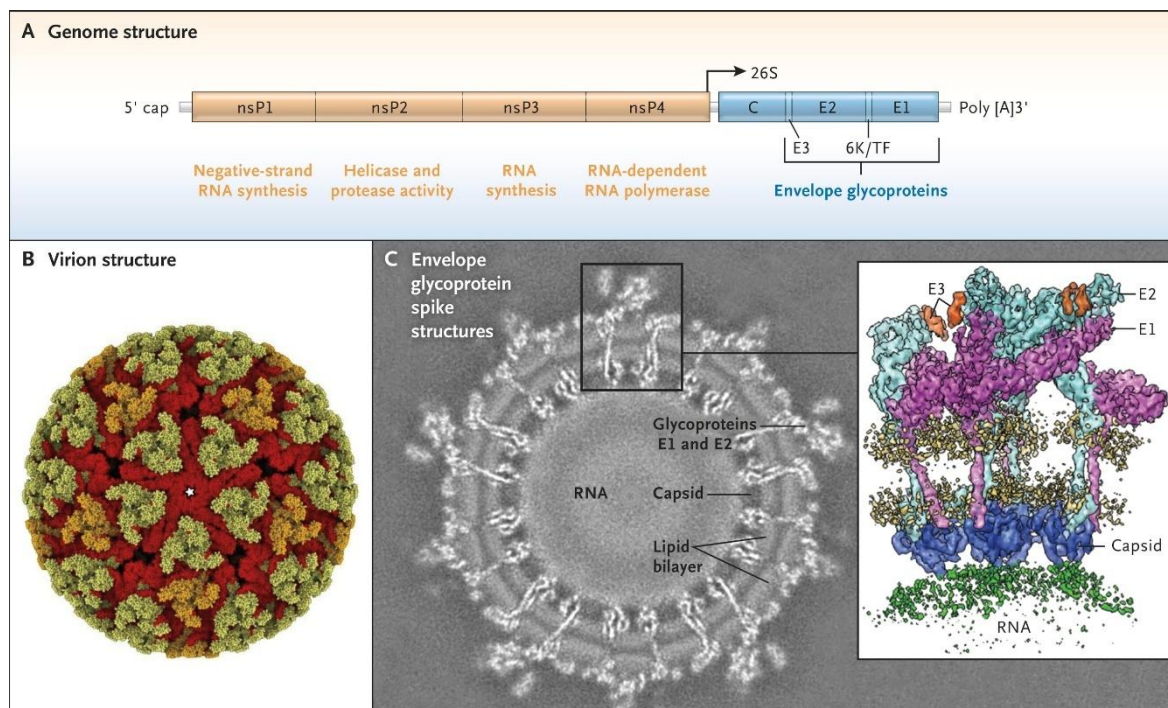


Figure 1.10. (A) CHIKV genome organisation with non-structural proteins (orange) and structural proteins (blue). (B) Virion structure Schematic taken from Felix Rey, Institut Pasteur, Centre National de la Recherche Scientifique. (C) Atomic resolution structures of spike protein (Voss et al., 2010) along with cryoelectron microscopic reconstructions of chikungunya (R. Zhang et al., 2011). Schematic taken from (Weaver & Lecuit, 2015).

Non-Structural Protein	Size (aa)	Function (Silva & Dermody, 2017)
nsP1	535	Methyltransferase and guanylyltransferase activity that caps viral RNA; sole membrane anchor
nsP2	798	N-terminal NTPase, helicase, and RNA triphosphatase activities; C-terminal cysteine protease activity responsible for processing of non-structural polyprotein
nsP3	530	Phosphoprotein with unknown functions, but important for minus-strand synthesis; contains macro domain and SH3-binding regions; likely interacts with host proteins
nsP4	611	RNA-dependent RNA polymerase (RdRp); putative terminal transferase activity

Table 1.4. The table is taken from Silva & Dermody, 2017 and outlines the functions of each non-structural protein of CHIKV, along with the corresponding number of amino acids for each nsP.

1.10. Replication cycle of Chikungunya virus

The CHIKV enters into the host cell mainly by clathrin-mediated endocytosis after binding to the plasma membrane receptor like MXRA8 (van Duijl-Richter et al., 2015). Upon attachment the viral genomic RNA (gRNA) is released into the host cell cytoplasm followed by E1-mediated fusion of the virion and the endosomal membrane. The RNA is then translated with host ribosome to produce the non-structural proteins as two polyproteins, nsP1-3 and nsP1-nsP4 (Frolov & Frolova, 2022). The nsPs are then separated with the help of the nsP2 protease activity (Strauss et al., 1983). Upon the translocation of the replicase complex the formation of replication spherules takes place. With the help of nsP4 RNA-dependent, RNA polymerase, negative-strand and new positive-strand viral gRNA and subgenomic (sgRNA) are produced (Ahola & Kääriäinen, 1995) (K. Zhang et al., 2022) (Laurent et al., 2022). The sgRNA is then translated to produce the structural polyprotein C-E3-E2-6K/TF-E1. Due to capsid autoproteolysis the free capsid is then released into the cytoplasm. The capsid then interacts with the genomic RNA and thus leads in the formation of icosahedral nucleocapsids (Holmes et al., 2020). The enveloped proteins are produced at the endoplasmic reticulum membrane and upon maturation are cleaved by trans-Golgi network-localized furin proteases and glycosylation (Jose et al., 2009) (Brown et al., 2020) (Figure 1.11.).

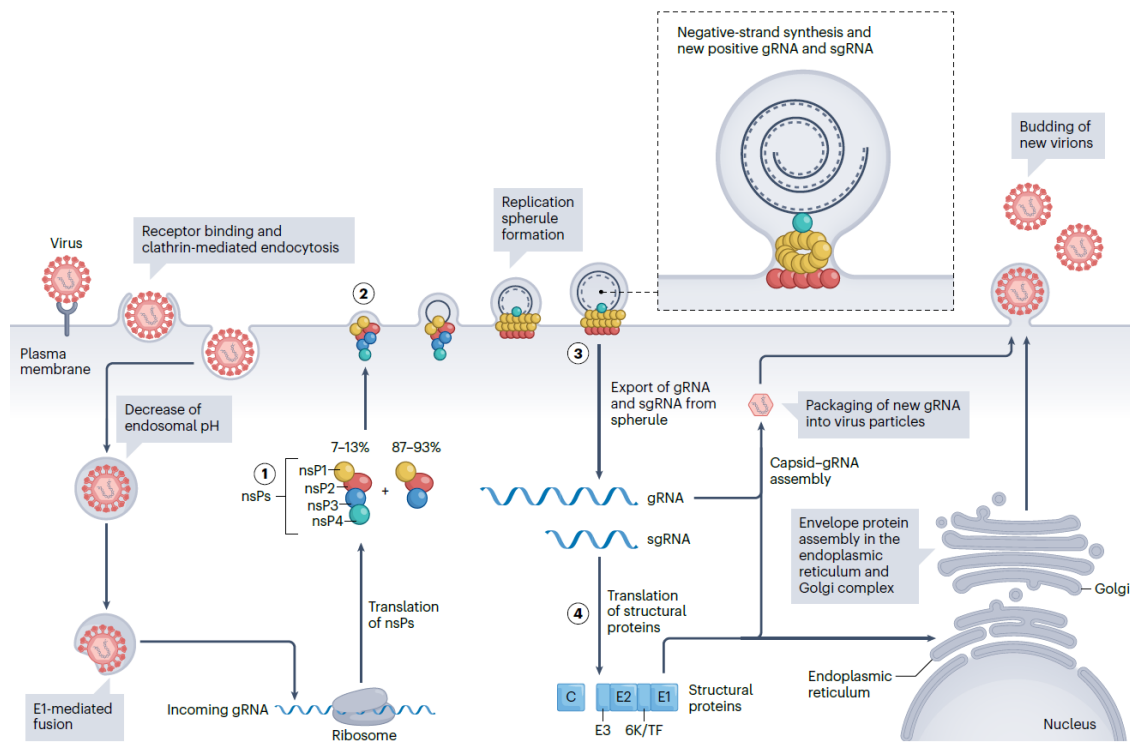
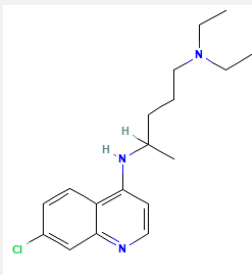
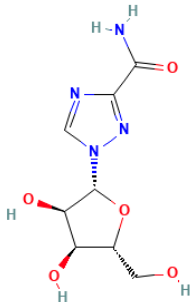
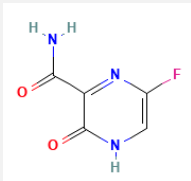
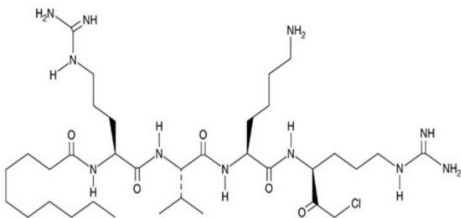
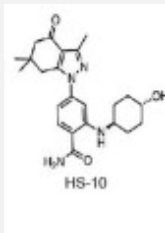
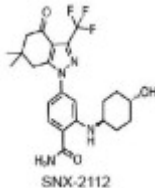


Figure 1.11. The viral replication cycle of CHIKV shows the binding of the virus to the plasma membrane resulting in clathrin-mediated endocytosis followed by a decrease of endosomal pH. Later the viral gRNA is translated by the host ribosome, resulting in nsPs, which is then followed by the formation of a replication spherule for the synthesis of negative strand and new positive gRNA as well as agRNA. The gRNA and sgRNA are released from the spherule for the translation of structural protein and capsid-gRNA assembly. Schematic taken from (Bartholomeeusen et al., 2023).

1.11. Treatment of Chikungunya virus infection

In order to combat the CHIKV infection many therapeutic strategies have been investigated which are briefly discussed by (Abdelnabi et al., 2015) and (Hucke & Bugert, 2020). While specific drug treatments for CHIKV infections is not known, there are repurposed drugs that have been used only for palliative care. The treatment of CHIKV infection includes rest, fluids, antipyretics and analgesics. Even though many antivirals and inhibitors have been identified against CHIKV infection, these are only tested on cell lines and have *in vitro* efficacy. Further pre-clinical and clinical testing is required to confirm their activities. Among the direct acting antivirals, chloroquine showed *in vitro* efficacy by inhibiting CHIKV infection in Vero A cells. The underlying mechanism of action is the inhibition of fusion of viral E1 protein with the endosomal membrane due to increase in the endosomal pH (Khan

et al., 2010) (Chopra et al., 2014). However, there was no significant *in vivo* efficacy in macaque model and in clinical trials. Another direct antiviral drug is Ribavirin which showed *in vitro* activity by inhibiting CHIKV replication in Vero cells ($EC_{50} = 341 \mu M$). It inhibits CHIKV genome replication via GTP depletion and also showed in-vivo efficacy by reducing viral load and inflammation in CHIKV infected mice when combined with doxycycline (Briolant et al., 2004) (Rothan et al., 2015). In a clinical study on ten patients with CHIKV infection and arthritis were given Ribavirin and it was found that there was improvement in the joint pains (Ravichandran & Manian, 2008). Favipiravir also falls under the direct-acting antiviral that inhibits replication of the viral genome. There is proven in-vitro evidence of inhibition of CHIKV induced cytopathogenic effect in Vero A cells ($EC_{50} = 25 \mu M$). The *in vivo* study indicated reduced mortality rate in the CHIKV infected AG129 mice (Delang et al., 2014). Another class of antivirals are host-targeting antivirals like Decanoyl-RVCR-chloromethyl ketone (dec-RVCR-cmk) which is an irreversible Furin peptide inhibitor. The cellular furins are responsible for processing E3E2 viral glycoprotein in order to produce mature virions. As the furins were inhibited by dec-RVCR-cmk, it resulted in the formation of immature virions thus lowering the viral spreading. The *in vitro* testing on human muscle satellite cells showed inhibition of CHIKV infection but the *in vivo* efficacy was not determined (Ozden et al., 2008). There are heat shock protein (HSP)-90 inhibitors like HS-10 and SNX-2112 which showed *in vitro* efficacy by reducing CHIKV infection in a dose dependent manner in HEK-293T cells. The *in vivo* efficacy was also determined, and it was found that viral titers were reduced in the serum of infected SvA129 mice upon treatment with these inhibitors (Rathore et al., 2014). There are also antivirals against the chikungunya virus infection with unidentified targets and mechanism of action, Aplysiatoxin is one such example. It inhibited CHIKV replication in SJCRH30 cells ($EC_{50} = 1.3 \mu M$) but its *in vivo* efficacy is not determined (Gupta et al., 2014) (Table 1.5.).

Antiviral Drug	Structure	Efficacy Testing
Chloroquine		Had in-vitro activity but no proven in-vivo efficacy.
Ribavirin		Had both in-vitro and in-vivo efficacy along with a clinical study.
Favipiravir		Had both in-vitro and in-vivo efficacy.
Decanoyl-RVCR-chloromethyl ketone (dec-RVCR-cmk)		Had in-vitro activity but no proven in-vivo efficacy.
HS-10		Had both in-vitro and in-vivo efficacy.
SNX-2112		Had both in-vitro and in-vivo efficacy.

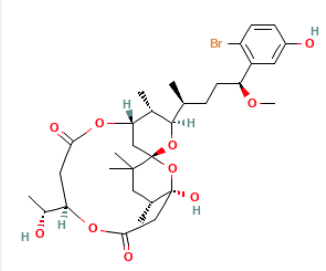
Aplysiatoxin		Had in-vitro activity but no proven in-vivo efficacy.
--------------	---	---

Table 1.5. CHIKV infection inhibitors with their structures and efficacy testing status.

Currently, there are no approved vaccines for humans against the CHIKV infection. But many vaccine candidates are currently in preclinical and clinical development. These are vaccines belonging to various platforms such as inactivated, live attenuated, live vectored, chimeric, virus-like particle (VLP), subunit protein and DNA vaccines (Weaver et al., 2012) (Smalley et al., 2016).

1.12. Bioinformatics- Multiple Sequence Alignment

Bioinformatics is essential for the management of data which is generated in the fields of modern biology and medicine. It helps in the analysis and interpretation of sequence data thereby enhancing discovery of drug targets (Bayat, 2002). The first step to understand the molecular pathway of an unknown sequence is to perform a sequence alignment. This is done by aligning the unknown sequence with one or more known sequences in order to identify similar or identical residues shared among them (Mount, 2004). The presence of these common residues suggests that they have a functional and structural role that is preserved by natural selection during evolution (Thompson et al., 2011). There are two types of sequence alignment; (a) pairwise sequence alignment (PSA); and (b) multiple sequence alignment (MSA). PSA is used for aligning two sequences whereas, MSA is used for aligning multiple sequences. There are two ways of aligning the sequences: first, local alignment, which only aligns a region within the sequence, and second, global alignment, which aligns the entire length of the sequence. MSA is more advantageous than PSA, as it can be used for the identification and quantification of conserved regions or functional motifs within a sequence family, thus pointing out possible evolutionary divergence between them (Kumar and Filipski, 2007). Sequence alignments also provide information regarding variations that are present in the sequences. Additionally, they provide insight into indels by incorporating gaps (Figure 1.12.).

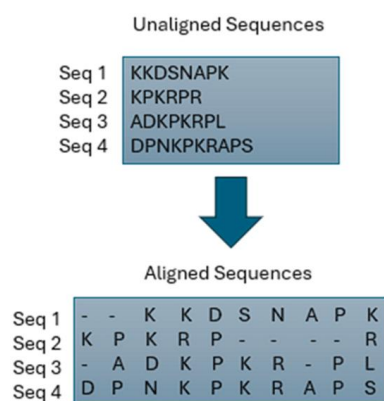


Figure 1.12. Example of a multiple sequence alignment of four protein sequences (Seq 1, Seq 2, Seq 3, Seq 4) showing unaligned as well as aligned sequences with gaps represented as (-). Schematic taken from (Ibrahim et al., 2024).

MSAs are widely utilised for sequence analysis in the field of molecular biology, computational biology, structure-based drug design and bioinformatics. With advances in computational power, different MSA algorithms have been developed to improve the accuracy of the alignments. There has been an exponential surge in the generation of sequence data with the aid of high throughput sequencing technologies. This negatively impacts the speed and accuracy of the alignment due to increased number of sequences. But now most recent MSA algorithms are capable of dealing with the increased number of sequences and longer lengths (Edgar, 2004). The computational algorithms used in MSA are based on either Dynamic Programming (DP) which is a slow, yet accurate method or heuristic/probabilistic method which is faster but less accurate (Daugelaite et al., 2013). DP is a mathematical and computational method used for dealing with complicated problems by subdividing them into smaller and much simpler components but in a repeated manner. This is utilised in global alignments by the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) and local alignments by using the Smith-Waterman algorithm (Smith & Waterman, 1981). These algorithms are generally more useful in performing PSA (two sequence alignments) and not for more than two sequences, as time and space complexity become enormous (Kato & Toh, 2008). Therefore, for generating MSAs, a heuristic approach is used to deal with the complexity involved due to presence of multiple sequences. The different types of PSA and MSA algorithms which are used for sequence alignments are mentioned in Table 1.6. There are different scoring methods which are used to denote the level of similarity or identity. For protein alignments,

the substitution matrices used are Point Accepted Mutation (PAM)(Y. Zhang et al., 2022), and Blocked Substitution Matrix (BLOSUM) (Henikoff & Henikoff, 1992).

Types of alignment	Algorithms
Pairwise Sequence Alignment	Needleman-Wunsch, k-mer, k-tuple, and Smith-Waterman algorithms.
Multiple Sequence Alignment	Progressive, Iterative, Heuristic

Table 1.6. Different types of sequence alignments and algorithms.

1.13. MSA Algorithms

The algorithms used in MSA are classified into progressive alignment, iterative, and heuristic methods Table 1.7.

Algorithm	MSA tools
Progressive Alignment	ClustalW, DiAlign, Kalign, T-Coffee
Iterative	MAFFT, MUSCLE
Heuristic	Hidden Markov Model (Clustal Omega), Genetic Algorithm

Table 1.7. Different types of MSA algorithms and their corresponding tools.

Progressive Alignment

Progressive alignment works by progressively building the full alignment, the first step is the completion of pairwise alignments using methods such as the Needleman-Wunsch algorithm, Smith-Waterman algorithm, k-tuple algorithm (Wilbur and Lipman, 1983), or k-mer algorithm (Edgar, 2004), and then the sequences are clustered together to show the similarity between them using methods such as mBed and k-means (Sievers et al., 2011). Later, the similarity scores are converted to distance scores followed by the construction of guide trees by guide tree building methods such as Neighbour-Joining (NJ) (Saitou and Nei, 1987) and Unweighted Pair Group Method with Arithmetic Mean UPGMA (Gronau and Moran, 2007). After the guide tree is built the MSA is assembled by the addition of the sequences into the alignment.

ClustalW (Thompson et al., 1994) is a MSA tool widely used in genome analysis, it first uses the k-tuple or Needleman-Wunsch method to perform pairwise sequence alignment. In the

k-tuple method similarity scores are estimated as the number of k-tuple matches (which are runs of identical residues, usually 1 or 2 for protein residues) in the alignment between pair of sequences. The similarity scores are thus generated which are converted to distance scores to further construct the guide tree. It uses weighted sum-of-pairs (WSP) objective function and the NJ method to build the guide tree as illustrated in Figure 1.13. The NJ method tracks the nodes on a tree rather than clusters of taxa. The similarity scores from the k-tuple method are stored in a matrix and a modified distance matrix is constructed considering separation between each pair of nodes. The guide tree is then generated by calculating an average value for divergence from all the nodes (Daugelaite et al., 2013).

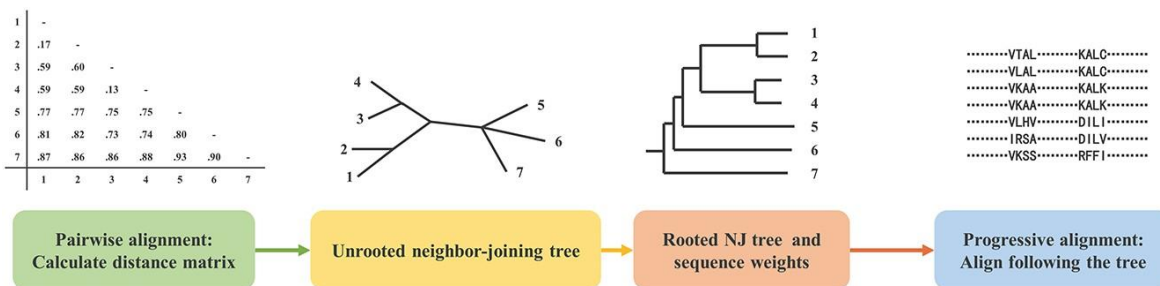


Figure 1.13.The workflow behind progressive alignment. First, the algorithm calculates a distance matrix of paired sequences (green box), followed by the construction of unrooted guide trees (yellow box) and then rooted guide trees with sequence weights (orange box). Finally, the sequences are aligned one by one (blue box).. Schematic taken from (Y. Zhang et al., 2022).

Kalign (Lassmann & Sonnhammer, 2005) uses a similar approach to perform MSAs as mentioned above. It uses progressive methods for sequence alignments like starting with calculating pairwise distances using the k-tuple method followed by the construction of guide tree using UPGMA or NJ method. It specifically uses the Wu-Manber approximate string-matching algorithm (Lassmann, 2020) (Wu & Manber, 1992) for distance calculation as well as in the dynamic programming which is used in aligning the sequences. This method allows string matching with mismatches (Daugelaite et al., 2013). The tool is capable of aligning large sequence datasets (discussed briefly in Chapter II).

T-Coffee (Notredame et al., 2000) known as Tree-based Consistency Objective Function For alignment Evaluation, uses a progressive alignment algorithm and consistency by assigning weights to each pair of aligned residues. The package also integrates many MSA methods

with different characteristics which aids in higher accuracy and low memory consumption (Y. Zhang et al., 2022).

Iterative algorithm

The iterative method performs continuous optimization of the alignment in order to improve the overall accuracy. It is divided into deterministic and stochastic algorithms (Notredame, 2002). The difference between the two algorithms is that the result of the deterministic algorithm remains the same each time whereas, the results of the stochastic algorithm may be different as it uses random values. In the iterative method the algorithm performs post-processing which involves making relevant changes by progressive methods thus modifying the construction of guide tree. The underlying mechanism involved here is the generation of new distance matrices using an MSA which is obtained by progressive alignment followed by the construction of a new guide tree. This new guide tree is used for second round of progressive alignment. There are more such rounds where the aligned sequences are divided into two groups followed by realignments until the alignment process is well converged (Gotoh, 1996) (Wang & Li, 2004).

MUSCLE, which stands for Multiple Sequence Comparison by Log-Expectation (Edgar, 2004), is based on an iterative algorithm to generate MSAs. There are three stages involved in the algorithm. The first stage involves the generation of a distance matrix (D1), which is used to generate a guide tree (TREE1) by UPGMA, followed by the generation of an alignment (MSA1). In the second stage, the Kimura distance matrix (D2) is generated using MSA1. The D2 matrix is then used to produce a second guide tree (TREE2), followed by the generation of another alignment (MSA2). In the last stage, the edge near the root node on TREE2 is deleted to generate two subtrees, followed by computing the subtree profiles (Rychlewski et al., 2000). These subtree profiles are then realigned to obtain a new MSA. Lastly, the sum-of-pairs (SP) scores are taken into account to consider if it is higher for the new MSA; if yes, the alignment is retained. This step is repeated until accuracy is achieved (Figure 1.14.).

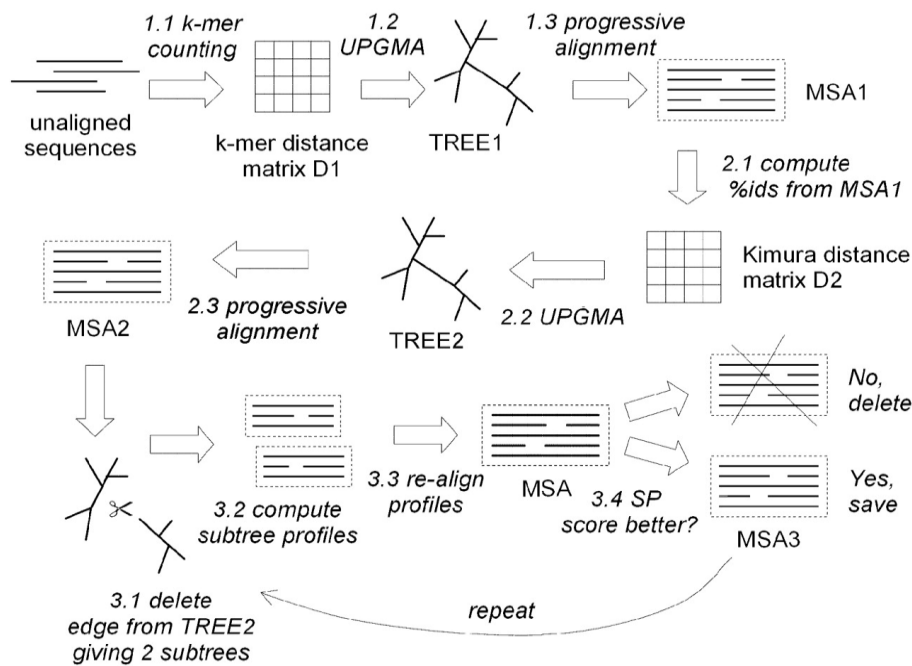


Figure 1.14. The stages involved in the MUSCLE algorithm. The first stage is the draft progressive stage, where the distance matrix (D1) is calculated using *k*-mer and UPGMA is used for building the guide tree (TREE 1), resulting in MSA 1. The second stage is the improved progressive stage, where the distance matrix is calculated using the Kimura distance matrix (D2), followed by the construction of the second guide tree (TREE 2), resulting in MSA 2. And the last stage is the refinement stage, where the edge from the TREE 2 is deleted, giving 2 subtrees; the profiles of the subtrees are computed, and the profiles are realigned, followed by observing the SP scores to see whether they are better or not, ultimately resulting in MSA 3. Schematic taken from Edgar, 2004.

MAFFT (Kato et al., 2002) is another iterative algorithm based MSA tool which stands for Multiple Alignment based on Fast Fourier Transform. MAFFT simplifies the calculations done by the algorithm as the input sequences are converted into a vector (volume and polarity values) for processing with the aid of Fourier transform. Due to this the homologous regions are quickly identified with quick turnover time. It uses two-cycle heuristics, the progressive method (FFT-NS-2) and iterative refinement method (FFT-NS-i). The pairwise distances are quickly calculated using FFT-NS-2 and a low quality MSA is produced which is used to calculate refined distances followed by one cycle of progressive alignment with FFT-NS-I (Daugelaite et al., 2013).

Heuristic algorithm

The heuristic algorithm basically utilises randomness and repeated steps to improve an alignment. It is known as a stochastic iterative algorithm which have good performance in solving hard problems quickly (Zhang et al., 2022). It is divided into genetic algorithm and Hidden Markov Model (HMM).

Hidden Markov Model (HMM)

HMM is a statistical analysis model used to represent a Markov process in which the parameters are not directly observable or known. It generates a probability matrix by utilising information like previous and current states of a test subject and thus predicts the next state (Baum et al., 1970). The MSA tools based on HMM are ProbCons and MSAProbs (Liu et al., 2010).

ClustalOmega (Sievers et al., 2011) is based on the mBed algorithm and HMM. It is capable of aligning large sequence datasets (e.g. 190,000) sequences on a single CPU processor within a few hours. In a benchmarking study, it was found that ClustalOmega produced MSAs that were more accurate and were obtained faster (Sievers et al., 2011). There is also a feature of reusing a pre-existing alignment and new sequences can also be added into this alignment. The MSA is produced by first producing pairwise alignments with k-tuple method followed by generating distance scores. The sequences are then clustered using the mBed and later by the k-means methods. The guide tree is produced using the UPGMA method and lastly the MSA is produced using the HAlign package, which performs progressive alignment using HMM (Figure 1.15.).

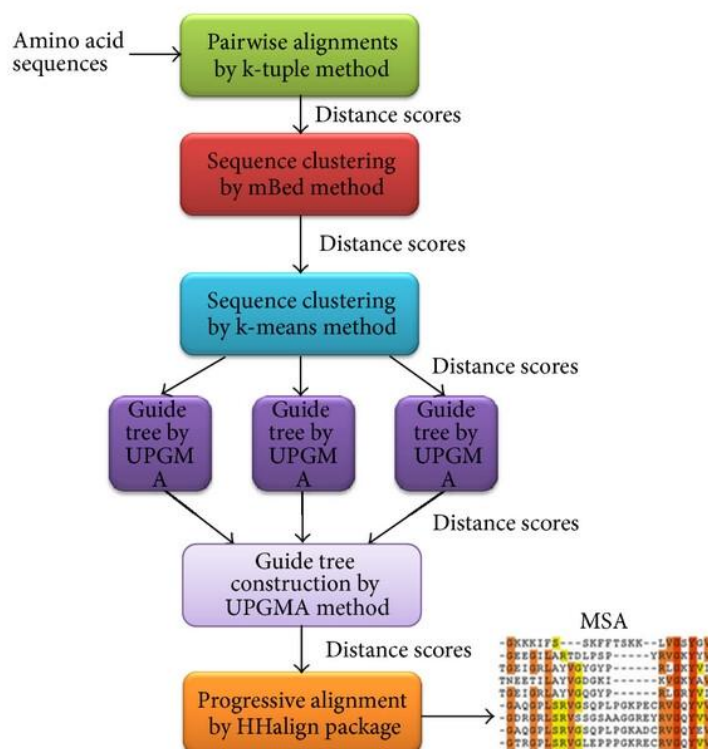


Figure 1.15. The workflow involved in Clustal Omega algorithm which produces MSA by first completing a pairwise alignment using k-tuple, sequence clustering using the mBed and k-means method, guide tree construction using UPGMA and finally progressive alignment using HHalign package. Schematic taken from (Daugelaite et al., 2013).

Alignment Scoring

The quality of the sequence alignment is measured quantitatively by the MSA tools which focusses on a scoring model. The most common type of scoring model used is sum-of-pair (SOP) where a pairwise matched residue will get a positive score, a mismatch and a gap gets a negative score. Considering MSAs, first a pairwise score is given to the sequence pairs which are summed to get the total SOP score. For, k numbers of sequences the total number of pairwise projections to be done are $k \times (k - 1)/2$ (Murata et al., 1985) so, the SOP equation can thus be defined as follows.

$$\text{SOP} = \sum_{i=1}^{k-1} \sum_{j=i+1}^k \delta(S_i, S_j) \quad (1)$$

Where S_i and S_j are two aligned sequences; $\delta(S_i, S_j)$ is the pairwise alignment score; k is the total number of sequences.

Gaps are inserted into the alignments when it deals with indels. These gaps are unlikely allowed to the hydrophobic groups of amino acids as they tend to fold to get inside the

protein core. On the other hand, gaps are tolerable to hydrophilic groups of amino acids (Chowdhury & Garai, 2017). Opening a gap is penalized more than extending gaps in an alignment. Once a gap is opened there is a higher chance that trailing gaps can occur. The penalty that is added is evaluated as follows.

$$W = \gamma + \delta \times (k - 1) \quad (2)$$

Where W is the total gap penalty, γ is the opening gap penalty, δ is gap extension penalty, and k is the total gap length (Gotoh, 1982).

Another method of scoring is consistency based scoring (Gotoh, 1990) which improves the early alignment stages by aligning an outgroup sequences during each pairwise alignment. For example, the pairwise alignment of sequences A and B and the pairwise alignment of sequences B and C will guide the alignment of sequences A and C. Lastly, probabilistic scoring method align sequences by using the likelihood (probability) that two residues occur and align together, rather than just fixed match/mismatch scores. They compare the joint probability of alignment with the background frequency of each residue to decide how good an alignment is (Chowdhury & Garai, 2017).

1.14. Computer-Aided Drug Design (CADD) techniques

Scientists all over the world aim to find cures for diseases in the form of “pills”. However, the very foundational knowledge of how these pills or drugs work sometimes remains a black box, considering the complexity of human protein networks. With the advances in computation, it is now possible to visualise the chemistry behind the interactions between drugs and human proteins, as well as design new drugs that target specific proteins, with the aid of various drug design software. This field of chemistry is known as rational Computer-Aided Drug Design (CADD). CADD has the capability of increasing the hit rate of novel drug compounds as it uses a more targeted approach than the traditional high-throughput screening (HTS) method. As a part of drug discovery, CADD is primarily used for: (1) filtering large molecule libraries into smaller sets of predicted actives that can further be experimentally validated often known as virtual screening (VS); (2) hit to lead optimization in order to improve the affinity or pharmacokinetic properties such as absorption, distribution, metabolism, excretion, and toxicity (ADMET); (3) designing novel molecules by joining fragments.

CADD approaches are classified into two broad categories: structure-based drug design (SBDD) and ligand-based drug design (LBDD). In the SBDD approach, the target protein structure is explored to calculate interaction energies for the predicted hits, whereas in LBDD the known active and inactive molecules are used for chemical similarity searches or construction of predictive, quantitative structure-activity relation (QSAR) models (Kalyaanamoorthy and Chen, 2011) (Figure 1.16.). As part of the SBDD method, if the structure of the target protein is not known, homology modelling or de novo structure prediction can be used. To perform virtual screening, molecular docking, and molecular dynamics (MD) simulation, machine learning methods can be used to identify virtual hits. On the other hand, for LBDD method, QSAR modelling, pharmacophore modelling, similarity search or machine learning are used to identify virtual hits.

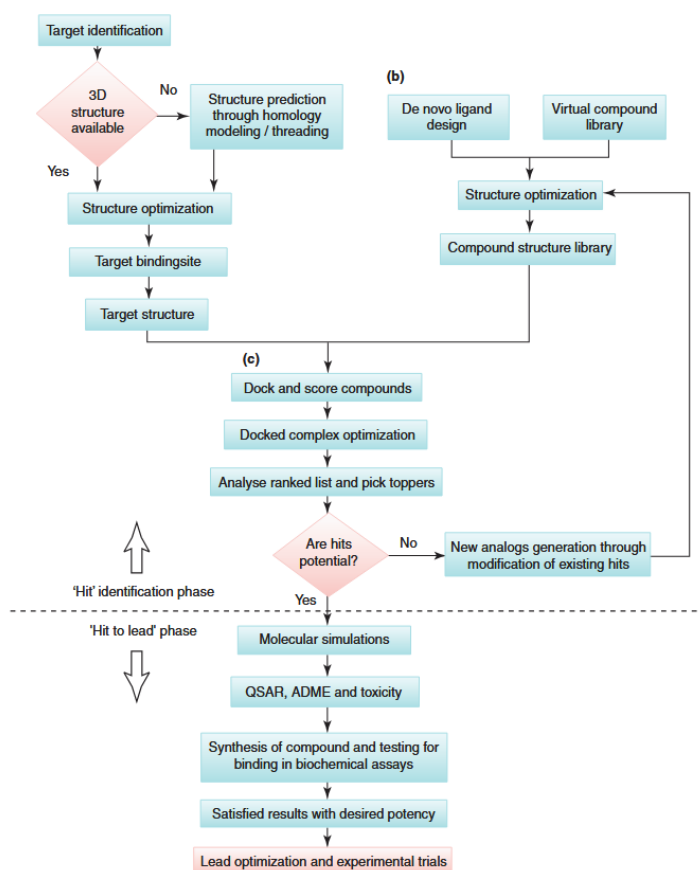


Figure 1.16. Classification of CADD into SBDD and LBDD methods to perform virtual screening for the identification of hits. The SBDD depends on the availability of 3D structure. If there are structures available, the structure will be optimised, and the binding site will be targeted for the identification of the molecule using docking and picking the top-ranked molecules. But if there are no structures available, LBDD is used, where a de novo ligand is designed, or a library is screened virtually. The ligand undergoes

structural optimisation, followed by docking with the target of interest. These are a part of the 'hit identification' phase. Upon identification of potential hits, these undergo further analysis, including molecular simulations, QSAR, ADME and toxicity identification. Interesting hits are synthesised and tested to identify potency; this phase is called the 'hit to lead' phase. Schematic taken from (Kalyaanamoorthy and Chen, 2011).

There were many successful studies that used CADD techniques to identify drug molecules for different protein targets. For example, the discovery of FDA-approved amprenavir, which was identified as the inhibitor of the human immunodeficiency virus (HIV) protease using SBDD techniques like protein modelling and MD simulations (Clark, 2006). Another example is the discovery of imatinib, which is a tyrosine kinase inhibitor used to treat chronic myeloid leukaemia (CML) (Roskoski et al., 2003). MD simulations were used to identify the interactions between imatinib and BCR-ABL, which in turn helped in constructing imatinib derivatives with better binding characteristics and selectivity for BCR-ABL (Ching et al., 2018). To treat glaucoma and macular oedema, dorzolamide was identified using fragment-based screening, which binds and inhibits the activity of carbonic anhydrase (Grover et al., 2006). A study used structure-based VS and pharmacophore modelling and identified isoniazid to treat tuberculosis by targeting InhA (Marrakchi et al., 2000). With the help of molecular docking, flurbiprofen was discovered, which binds to cyclooxygenase-2 to treat rheumatoid arthritis and osteoarthritis (Dadashpour et al., 2015). In our study we have used three techniques: (1) fastROCS; (2) pharmacophore-based virtual screening; and (3) docking.

FastROCS

FastROCS is an LBDD VS tool developed by OpenEye tool which uses fast shape comparison to screen molecules in just seconds that have similar shape to that of the query molecule (Figure 1.17.) (Hawkins et al., 2007). If their volumes overlay well, it outputs combination of similarity scores (Tanimoto-shape and colour) and whereas any volume mismatch is a measure of dissimilarity (OpenEye Documentation) (discussed briefly in Chapter III).

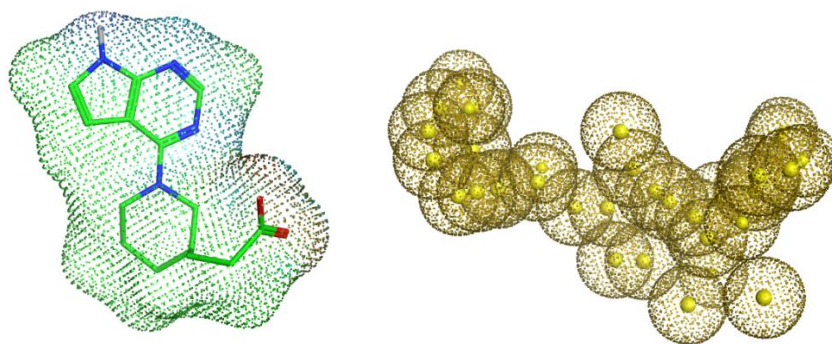


Figure 1.17. Examples of volume queries of ligand (green) and dummy atoms (yellow) used as inputs into the fastROCS volume search. The picture was generated with MOE (2022).

Pharmacophores

Pharmacophores are a powerful tool to describe and visualize ligand-target binding models. The official IUPAC definition of pharmacophores (Wermuth et al., 1998) states that ‘A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supra-molecular interactions with a specific biological target structure and to trigger (or to block) its biological response’. The pharmacophore features are 3D geometric representations such as spheres, planes or vectors that describes the nature and location of the chemical features (Table 1.8.) that are necessary for ligand-target interactions (Seidel et al., 2020). Pharmacophore models can be generated manually or automatically by taking in consideration either the structure of known active ligands (ligand based) (Figure 1.18.) or from the 3D structure of the target protein (structure based) (discussed briefly in Chapter III).

Feature Type	Geometric Representation	Complementary Feature Type(s)	Interaction Type(s)	Structural Examples
Hydrogen-Bond Acceptor (HBA)	Vector or Sphere	HBD	Hydrogen-Bonding	Amines, Carboxylates, Ketones, Alcohols, Fluorine Substituents
Hydrogen-Bond Donor (HBD)	Vector or Sphere	HBA	Hydrogen-Bonding	Amines, Amides, Alcohols
Aromatic (AR)	Plane or Sphere	AR, PI	π -Stacking, Cation- π	Any aromatic ring
Positive Ionizable (PI)	Sphere	AR, NI	Ionic, Cation- π	Ammonium Ion, Metal Cations
Negative Ionizable (NI)	Sphere	PI	Ionic	Carboxylates
Hydrophobic (H)	Sphere	H	Hydrophobic Contact	Halogen Substituents, Alkyl Groups, Alicycles, weakly or non-polar aromatic rings

Table 1.8. The summary of the most important pharmacophore features, their geometric representations, complementary feature types, structural examples and their interactions with the target. Adapted from (Seidel et al., 2020).

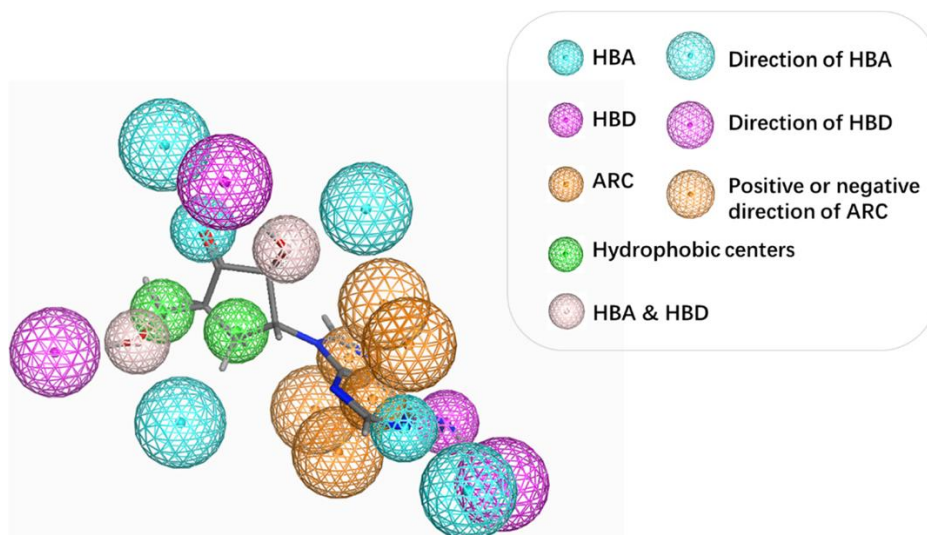


Figure 1.18. Example of pharmacophore features describing the properties of the ligand. Cyan mesh represents hydrogen bond acceptor (HBA), pink mesh represents hydrogen bond donor (HBD), orange mesh represents aromatic centre (ARC), green mesh represents hydrophobic centres, and light pink mesh represents both HBA and HBD. Image reference (Wang et al., 2020)

Molecular Docking and protein-ligand interactions

Molecular docking is used in CADD for predicting the binding modes and affinities of small molecules against their targets that are usually proteins (Meng et al., 2011). The small molecules can be docked in four ways, i.e., (a) by rigid docking, where both target and ligands are treated as static entities; (b) by flexible docking, where both ligand and target are considered to be flexible; (c) by flexible ligand docking, where the ligand is considered to be flexible and the target is considered static; and (d) by flexible protein docking, where the target is considered to be flexible and the ligand is considered static (Mohan et al., 2005). The aim of molecular docking is to identify ligands that bind to the binding pocket present within a receptor and to determine its most energetically stable binding poses. Binding poses can be described as the conformation of the ligand within the binding pocket that results in the ligand-protein interactions. The scoring function estimates the binding affinities of the generated poses and ranks them followed by identification of the most

favourable ligand-protein binding modes (Figure 1.19.) (Du et al., 2016). There are different types of non-covalent interactions present between the ligand and the protein (Figure 1.20.). Among them hydrogen bonds are the second most frequent type of ligand-protein interactions. The hydrogen bond is a result of attractive interaction between a hydrogen atom that is attached to a more electronegative (donor) atom and an acceptor atom having electrons available for sharing (Daze and Hof, 2016) depicted as $D-H\cdots A$ (Figure 1.21.). The molecular docking approach is widely used in SBDD virtual screening process (discussed briefly in Chapter III).

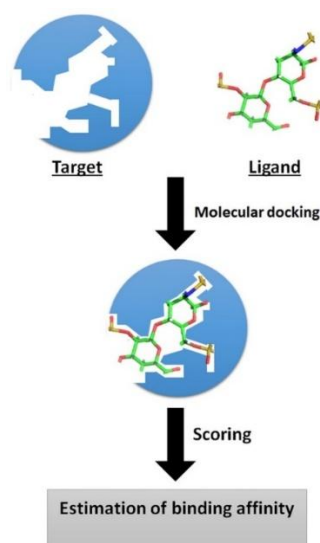


Figure 1.19. Principle of molecular docking where the target binding site is represented in blue, and the ligand is represented in green. Image reference (Baig et al., 2018)

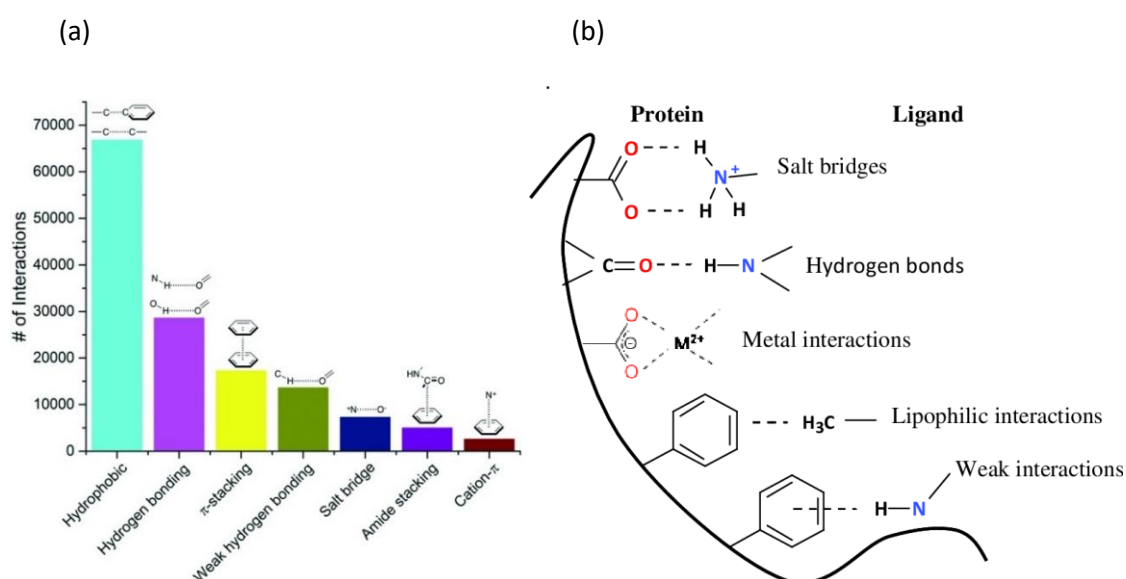


Figure 1.20. Different types of non-covalent interactions between protein-ligands with their (a) frequency distributions (Ferreira de Freitas and Schapira, 2017) as well as how the (b) the ligand interacts with protein residues with the presence of salt bridges, hydrogen bonds, metal interactions, lipophilic interactions and weak interactions. (Haider, 2010).

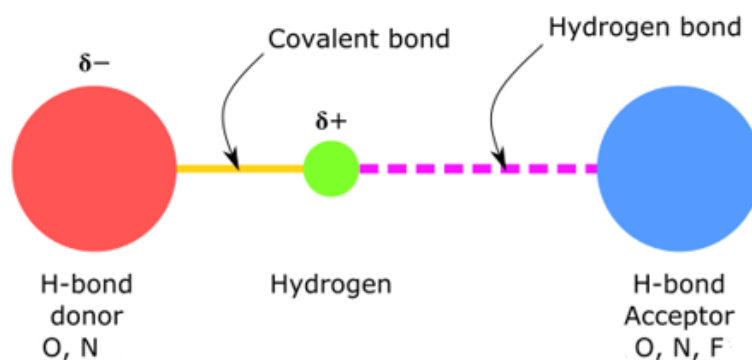


Figure 1.21. Hydrogen bond (represented by dotted line) formation between a hydrogen atom covalently bonded to a electronegative atom like (O-oxygen, N-nitrogen) and a nearby acceptor atom (O-oxygen, N-nitrogen, F-Florine). Schematic taken from <https://chemistrygod.com/hydrogen-bonding-chemistry121>.

Molecular Dynamics (MD) and Absolute Binding Free Energy (ABFE)

The molecular docking method results in a protein-ligand complex pose, which is static. In order to understand and explore the interactions between the protein-ligand complex, it is essential that the dynamic behaviour is studied. Proteins are highly dynamic in nature and tend to acquire multiple conformations; hence, it is important to understand what conformation the protein will acquire upon binding to the ligand. To solve this, MD simulations are carried out. MD simulations model the motion of each atom in a protein or molecular system over time by applying physical principles that describe how atoms interact with one another (Karplus & McCammon, 2002). There are two types of simulation an all-atom based simulation and coarse-grained simulation. In all-atom simulations every individual atom in a system is modelled for high-detail accuracy, while in coarse-grained simulations groups of atoms are combined into larger particles to enable faster, large-scale or long-timescale modelling. The MD setup depicts a real-life scenario of protein-ligand complex present within a solvent box (Figure 1.22.) (discussed briefly Chapter IV).

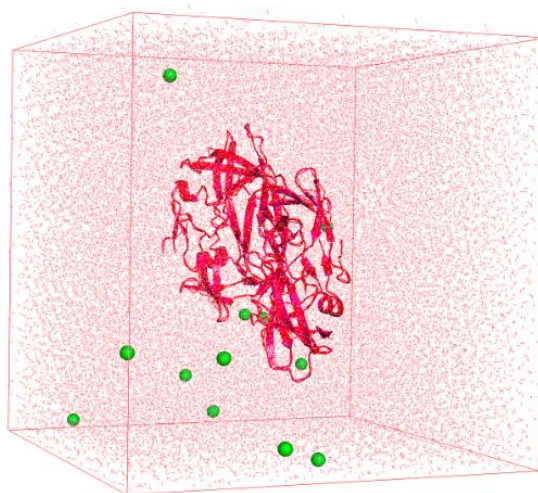


Figure 1.22. A typical MD setup with protein (magenta), ions (green) and solvent (red dots) in a cubic solvent box. The ions are added to balance the overall charge of the system. Schematic taken from Setting up a Molecular Dynamics simulation – Compchems

To model a protein-ligand interaction force fields (FFs) (Table 1.9.) are used because they provide the mathematical framework to calculate the forces and energies acting on atoms and molecules. They are used to describe both intermolecular and intramolecular interactions. After computing the forces acting on individual atoms the atom positions are updated by calculating acceleration and velocities using classical Newton's law of motion (Hospital et al., 2015). Different software packages that are used to run MD simulations include GROMACS (Van Der Spoel et al., 2005), NAMD (Phillips et al., 2005), OpenMM (Eastman et al., 2024), CHARMM (Brooks et al., 2009) and AMBER (Case et al., 2005).

No.	Forcefield	Drugs	Lipid	DNA & RNA	Protein
1	OPLS	OPLS-AA	OPLS-AA	OPLS-AA/M	OPLS-AA, OPLS-AA/L
2	GROMOS	GROMOS 45A3/4, GROMOS53A5/6, GROMOS54A8	GROMOS 43A1, GROMOS54B7	GROMOS53A6, GROMOS54A8	GROMOS54A7, GROMOS 43A1, GROMOS 45A4,

3	CHARMM	CHARMM general force field (CGenFF)	CHARMM27 lipids, CHARMM36 lipids	CHARMM27 DNA, CHARMM27 RNA/DNA, CHARMM 36 RNA, CHARMM 36 DNA	CHARMM22/CMAP, CHARM27, CHARMM36, CHARMM36m
4	AMBER	General AMBER force field (GAFF)	LIPID14, LIPID21	AMBER99 OL3, AMBER99bsc, AMBER OL15	AMBER94, AMBER96, AMBER99, AMBER99sb, AMBER03, AMBER14sb, AMBER15ipq, AMBER19sb
5	MARTINI	MARTINI 2, MARTINI22, MARTINI22p, MARTINI 3, MARTINI dry, MARTINI ELNEDYN22, MARTINI ELNEDYNP22	MARTINI 2, MARTINI22, MARTINI22p, MARTINI 3, MARTINI-Dry, MARTINI ELNEDYN22, MARTINI ELNEDYNP22	MARTINI 2015	MARTINI 2, MARTINI22, MARTINI22p, MARTINI 3, MARTINI dry, MARTINI ELNEDYN22, MARTINI ELNEDYNP22

Table 1.9. List of forcefields to model biomolecule-ligand interactions for MD simulations. Adapted from (Sinha et al., 2022).

The ultimate goal of an *in silico* hit finding approach is to accurately predict a binding free energy value, which represents how much the ligand (chemical entity) has an affinity for the target protein. In simple terms, binding affinity represents the value of how likely it is that the ligand will bind to the protein. The binding affinity represents the Gibbs free energy value (ΔG_{Bind}) for a protein-ligand complex (Figure 1.23.). The absolute binding free energy

is an “alchemical” method to predict binding affinity between the ligand and the protein using non-physical intermediate states by modifying interactions in the system (Mey et al., 2020). ABFE utilises MD simulations to carry out non-physical transformation in bound (complex in solvent) and unbound states (ligand in solvent) (discussed briefly Chapter IV).

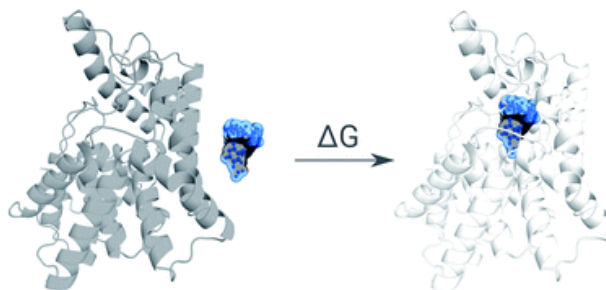


Figure 1.23. ABFE is a protocol which helps in the prediction of ΔG_{Bind} which is the binding affinity of the ligand (blue) with the target protein (grey and white). Schematic taken from (Khalak et al., 2021).

The drug discovery process has been accelerated with the help of CADD. Used appropriately, it can decrease timelines and is a cost-effective approach for identifying novel hit molecules and for improving a molecule's drug-like properties by pharmaceutical companies. VS of possible inhibitors against target proteins can give insight into the mechanism of action (MOA) of the molecule by analysing the binding interactions between the two.

Looking at the urgency of discovering drug treatments for COVID-19, repurposing of existing drugs was considered as a possible approach for finding quick therapeutic alternatives. The popular protein targets of SC-2 have been the proteases and the polymerase. Using docking based virtual screening (DBVS) Pant et al. (2020) calculated the binding affinity of FDA approved protease inhibitors against Mpro of SC-2. After screening large databases such as ZINC/ChEMBL, drugs such as cobicistat, ritonavir, lopinavir and darunavir were found to be the top hit molecules but were not confirmed *in vitro*. Subsequent biological evaluation of these compounds by other groups highlighted variation in prediction success. An *in vitro* study by Shytaj et al. (2022) found that cobicistat does not inhibit the enzymatic activity of Mpro but has an effect on S-glycoprotein fusion with an IC_{50} of 3.8 μM . The drugs such as lopinavir and ritonavir had an IC_{50} value of 12.01 μM and 19.88 μM respectively against SC-2 infection in Vero E6 cells (Zhang et al.,

2020). On the other hand, darunavir showed no antiviral activity against SC-2 at clinically relevant concentrations ($EC_{50} > 100 \mu M$) (De Meyer et al., 2020).

Considering protease inhibitors, a study by Wu et al. (2020) reported that many drugs displaying diverse therapeutic actions had predicted high binding affinity to PL^{pro} (NSP3). These drugs were a series of anti-viral drugs (ribavirin, valganciclovir and thymidine), anti-bacterial drugs (chloramphenicol, cefamandole and tigecycline), a muscle relaxant drug (chlorphenesin carbamate) and an anti-tussive drug (levodropropizine). *In silico* studies have suggested that ribavirin has a broad-spectrum impact on SC-2, acting at different viral proteins. It has been reported to show antiviral activity in Vero E6 cells following SC-2 infection as it decreases the expression of TMPRSS2 at both mRNA and protein levels 48 h after treatment (Unal et al., 2021), while promising does not necessarily mean that the MOA is via PL^{pro}. The RdRp inhibitors remdesivir and favipiravir have shown efficacy against SC-2. These were also found to be amongst high-ranking nucleotide analogues that could tightly bind to RdRp (Jang et al., 2021).

The NSPs are now emerging as a new avenue of investigation for identifying possible inhibitors that could block protein activity thus stopping viral replication. A study by Tazikeh-Lemeski et al. (2020) used a DBVS approach to screen FDA approved drugs against NSP16 to predict their inhibitory activity. The study predicted that anti-viral drugs (maraviroc and raltegravir) and an anti-inflammatory drug (prednisolone) could be effective drug candidates against NSP16 but were not confirmed *in vitro*. Our previous study (Kandwal and Fayne, 2022) that focussed on elucidating possible mechanisms of action of compounds that were active against whole virus SC-2, targeted NSPs of SC-2 and resulted in target prediction of many inhibitors based on pharmacophore based virtual screening. For example, the pharmacophore features selected for NSP15 resulted in AT-9283, acadesine, olomoucine, sapropterin and tetrahydrofolic acid showing promising interactions in the NSP15 binding site. Computational drug discovery-based approaches can help in discovering small molecule inhibitors of SC-2 viral proteins but there is also a need to validate these hits with *in vitro* and *in vivo* assays.

The structural proteins, such as the spike protein, can also be good targets. Many drugs such as anti-hypertensive drugs (rescinnamine), anti-fungal drugs (posaconazole), anti-

bacterial drug (sulfasalazine) and anti-coagulant drug (dabigatran etexilate) showed high predicted binding affinity for the spike protein of SC-2 (Wu et al., 2020). Among these drugs posaconazole specifically blocks SC-2 entry, possibly by selectively targeting the “E-L-L” motif of the S protein. The IC_{50} of the drug was estimated to be 3.37 μ M in HEK293T-ACE2 cells (Jana et al., 2022). An interesting study by Lau et al. (2021) found top scoring compounds (docking calculation) such as Imatinib, lapatinib, adapalene and candesartan cilexetil predicted to inhibit the binding domain of spike 1 (S1 subunit) of SC-2 and was followed up with an *in vitro* cellular infection assay. Among them imatinib and lapatinib had IC_{50} values of approximately 10 μ M.

1.15. Hypothesis and Objectives

The main hypothesis of this project is that there are conserved binding pockets in the NSPs of SC-2 that can be targeted by identifying small molecule inhibitors using CADD methods. It is also hypothesised that these conserved binding pockets might also be present across the cousin viruses of SC-2, like the Beta coronaviruses. And not only that, these conserved binding pockets might also be present in Group IV viruses. The idea is to identify a pan-antiviral inhibitor.

The objectives of the project are as follows:

- Perform MSA on large datasets of sequences to identify the level of amino acid conservation within the binding pockets of SC-2 NSP's.
- Visualise the conserved amino acid residues on a 3D-protein structure.
- Perform a tiered virtual screening protocol (fastROCS, pharmacophore search, and docking) to target the conserved binding pockets by screening molecules from the Enamine database.
- Develop an MD and ABFE simulation protocol to support experimentally determined activity values.
- Expand the study to look for conservation of these binding pockets across beta-CoVs and group IV viruses (CHIKV).

1.16. Opportunities

In June 2022, the CACHE challenge (<https://cache-challenge.org/>) was announced to discover inhibitors targeting the RNA-binding tunnel of SC-2 NSP13 (Challenge #2). By that time, we had already investigated the conserved residues present within NSP13 and had identified NSP13 as a protein of interest so decided to participate in the challenge. The cheminformatics phase of the project was scheduled to be started in early 2023 but was started in 2022 instead due to our participation in the challenge in which we were ranked among the top 3 globally. Later on, in December 2022 another challenge was announced by the CACHE organisation to discover inhibitors that target the Mac1 domain of NSP3 (#Challenge3). As our analysis for NSP3 conservation was also completed, we decided to participate in this challenge as well and were ranked among the top 4 globally. The predicted compounds are procured and tested at CACHE using both enzymatic and binding assays. The cost of chemical synthesis and biological assays were fully covered by CACHE for NSP13 challenge but for NSP3 challenge we covered 50% of the cost for chemical synthesis and rest was covered by CACHE including the biological testing. The experimental validation results for NSP13 and NSP3 are discussed in Chapter III.

Chapter II.

Bioinformatics

Parts of this chapter have been published in:

Kandwal S, Fayne D. Genetic conservation across SARS-CoV-2 non-structural proteins - Insights into possible targets for treatment of future viral outbreaks. *Virology*. **2023** Apr;581:97-115. doi: 10.1016/j.virol.2023.02.011. Epub 2023 Mar 10. PMID: 36940641; PMCID: PMC9999249.

Kandwal, S., Cmelo, I., & Fayne, D. SARS-CoV-2 ORF8 sequence conservation and mutational analysis—Insight into the influence of dataset size on identifying top mutations. *bioRxiv*, **2025.06.16.659836**. (pre-print, submitted for peer-review).

2.1. Introduction

This chapter will cover the results derived from the bioinformatics study conducted with the MSA tool Kalign, which was utilised for generating sequence alignments on SC-2 NSP protein sequences to identify conserved binding sites.

2.2. Kalign

Kalign (Lassmann & Sonnhammer, 2005) is a MSA tool which is used for producing high quality alignments of large volume of sequence datasets in an accurate and feasible way. The Kalign algorithm is based on progressive alignment which uses Wu-Manber approximate string-matching algorithm (Wu & Manber, 1992) for fast and accurate distance estimation. In sequence alignment similar looking sequences are aligned first, this is done to avoid errors at the very beginning due to the alignment of dissimilar sequences. To avoid this distance matrix and guide tree are generated. The first step of the Kalign algorithm is to perform pairwise alignments and generate matrix of pairwise distances between all the sequences. In Kalign, Wu-Manber approximate string-matching algorithm is used for calculating the pairwise distances based on the Levenshtein edit distance. The algorithm effectively searches short three-residue patterns with a time complexity of $O(tk)$, where t is the length of the text string and k is the number of errors allowed. The Wu-Manber algorithm is capable of detecting similarity in divergent or low identity sequences (below 30% identity) unlike exact similarity k-tuple methods that fails to detect identity (Lassmann & Sonnhammer, 2005). The overall accuracy of the alignment is improved by scoring matched patterns based on match quality and diagonal locality in a dynamic programming matrix (16 for exact matches, 8 for one mismatch, 1 for two mismatches). This defines sequence similarity as the sum of the top three diagonal scores which reduces the impact of spurious matches and provides fast, accurate similarity estimates. The matrix is then used to build the guide tree with UPGMA clustering method for progressive alignment. In the progressive alignment step, Kalign aligns sequences at each internal node of a guide tree. In the case of insertions or deletions it can optionally use Wu-Manber pattern matches as anchor points during alignment in order to improve accuracy. The alignment itself is carried out using global dynamic programming with affine gap penalties (Durbin et al., 1998), based on a model that prevents consecutive gaps. It uses one-dimensional arrays and a trace matrix for alignment reconstruction, reducing memory and

improving speed. If anchor points are included, there are two additional steps which are performed. First (consistency check), to check the reliability of the pattern matches without gap penalties and second (updating of pattern match positions), to track pattern positions as gaps are introduced in profiles. For the scoring, Kalign uses single soft substitution matrix GONNET250 (Gonnet et al., 1992) as it performed well considering similar and distantly related sequences without the use of dynamic matrix selection.

The old Kalign version used unweighted pair group method with arithmetic mean (UPGMA) algorithm to construct a guide tree which resulted in quadratic time complexity when dealing with large sequence datasets. Hence a new version of Kalign (Lassman, 2020) was released which was based on using SIMD (single instruction, multiple data) accelerated version of Gene Myers' bit-parallel algorithm (Myers, 1999) for generating the pairwise sequence distances. To speed up the construction of the guide trees it incorporated embedding strategy proposed by Blackshields et al., 2010 (Figure 2.1.).

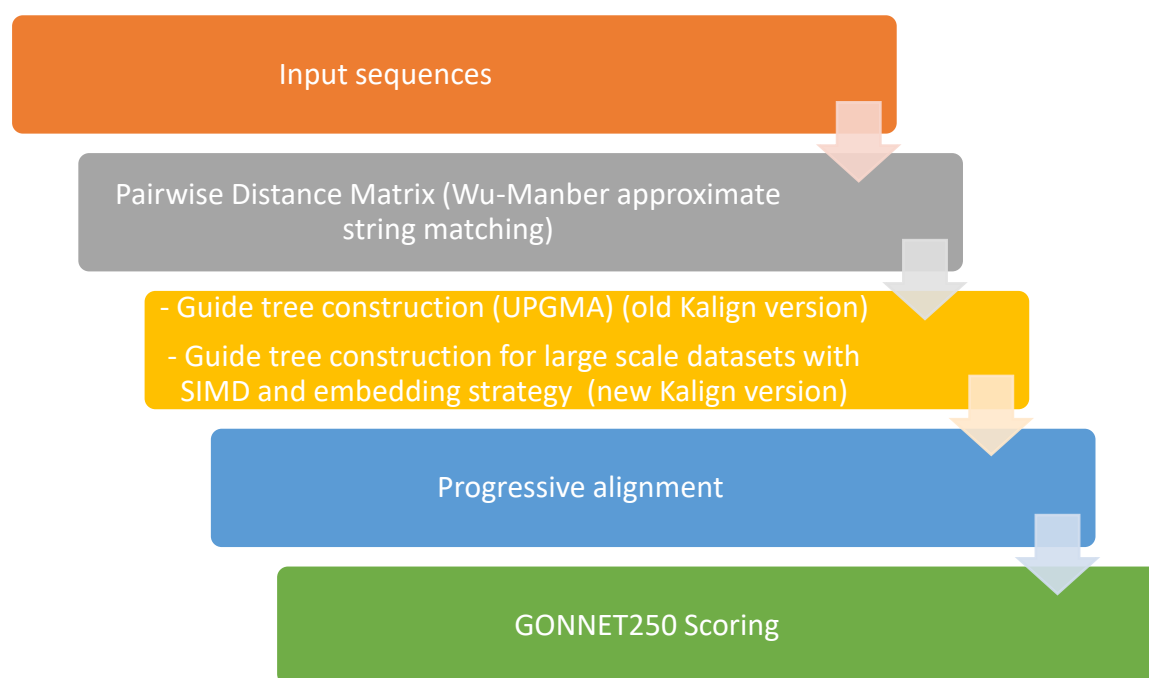


Figure 2.1. Kalign algorithm flow chart. Upon the introduction of the input sequences, the pairwise distance matrix is generated, followed by the construction of a guide tree and progressive alignment, which is scored using GONNET250.

The benchmarking study conducted by Lassman et al., 2020 on the Balifam dataset showed that Kalign outperforms Clustal Omega and MUSCLE sequence alignment tools in terms of computational speed (Figure 2.2.).

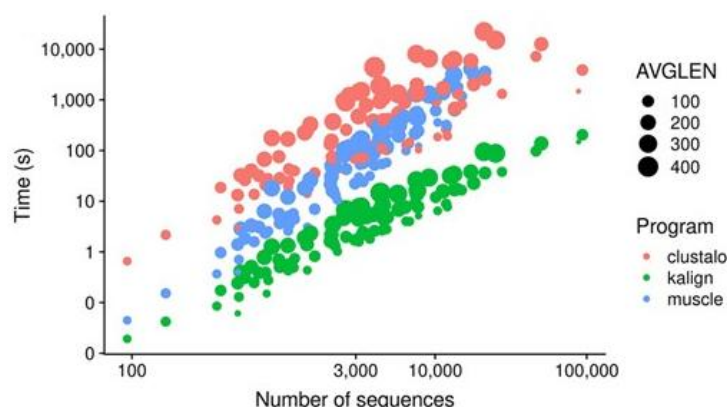


Figure 2.2. Computational performance of Kalign (green) against Clustal Omega (orange) and MUSCLE (blue). The x-axis denotes time (s) and y-axis denotes number of sequences. Schematic taken from (Lassman et al., 2020).

2.3. Conserved regions

Conserved sequences are either amino acid or nucleotide sequences that remain relatively unchanged far back up the phylogenetic tree or among diverse species. Sequences considered to be highly conserved suggest that the sequence has been maintained by natural selection. The conserved regions present within the sequences can be identified using multiple sequence alignment (MSA) studies and suggest structural or functional importance to the protein. To outline the importance of sequence and structural conservation across viruses, we have first focused on conservation between different RNA viruses followed by conservation within coronaviruses and then narrowed our focus to conservation of NSPs across coronaviruses. The highly conserved regions, if present in the binding pocket of the protein, can be interesting targets for developing small molecule inhibitors.

2.4. Similarities of constituent proteins within RNA viruses

In this section we focus on similarities across different RNA viruses. The RNA-dependent RNA polymerase (RdRp) (NSP12), which is essential for RNA replication, is known to be conserved within RNA viruses so targeting this protein may be an effective therapeutic approach (Xu et al., 2003; Gerlach et al., 2015). A study by Mönttinen et al. (2021) used 42

high-resolution viral RdRp structures from the PDB of -ssRNA, +ssRNA and dsRNA viruses. By using the sequence alignment tool MAFFT L-INS-i, amino acid alignment of complete RdRp sequences was performed. The catalytic site failed to align but 231 residues of the common structural core of the polymerases covered over 30% of the influenza A virus, over 50% of the poliovirus and 25% of SC-2 polymerase structures. The study pointed out possible structural conservation across the viral RdRps thus indicating the potential to target the polymerase for discovering wide spectrum antivirals.

Considering NSP3, the first 160 residues of the N-terminal region, also known as the macro domain (X domains), was found to be conserved among alphavirus, rubivirus and coronavirus in a sequence similarity study (Gorbalenya et al., 1991). Torovirus (ToV) belongs to the order *Nidovirales* and historically belonged to the family *Coronaviridae*. It now belongs to separate family *Tobamviridae* (ICTV, 2019). The six NSP domains shared by the ToVs and CoVs are 3C-like protease (NSP5), nidovirus RdRp-associated nucleotidyltransferase (NiRAN) of RdRp (NSP12), zinc-binding domain (ZBD) of helicase (HEL) (NSP13), and endoribonuclease (EndoU) (NSP15) which are known to be conserved in most nidoviruses (Van Boheemen et al., 2012; Saberi et al., 2018). The structural or sequence conservation among these viruses could support the hypothesis of having a common ancestor. These studies could help discover new antivirals targeting these conserved regions among diverse ranges of viruses.

The hepatitis C virus non-structural protein (NS5b) RdRp and MERS-CoV RdRp share a low sequence similarity but have structural similarity as both have two conserved aspartate residues at the active site (Elfiky and Azzam, 2021). The RdRp of flaviviruses such as Zika, West Nile Virus and Dengue virus share 66–69% amino acid sequence identity with the RdRp of SC-1 (Xu et al., 2003). Another approach to look for similarity between viruses is to analyse the structural similarity between them. In a study by Bafna et al. (2020), the PDB was searched using the *DALI* program (Holm and Sander, 1993, 1999) for 3D structural similarity with domains I and II (excluding domain III) of the SC-2 M^{pro} as the query search resulted in several structurally similar proteases, among them the hepatitis C virus NS3/4A serine protease. The proteases of HCV and SC-2 had a structural similarity Z score of +8.4 with an overall backbone root-mean-square deviation for structurally similar regions of approximately 3.1 Å. The superimposition of the two proteases resulted in overlap of their substrate binding clefts as well as their active-site catalytic residues, His41/Cys145 of the

SC-2 M^{pro} cysteine protease and His57/Ser139 of the HCV NS3/4A serine protease. In light of these structural similarities, HCV protease inhibitors were predicted to bind to the substrate-binding cleft of SC-2 M^{pro} and inhibit virus replication. The results of the study suggested that ten HCV protease inhibitors had the potential to bind with the M^{pro} binding cleft with hydrogen bonding and hydrophobic contacts. After biological testing, seven of the HCV drugs inhibited SC-2 replication in human 293T cells with IC₅₀ values from 0.55 to 20.5 μ M (Bafna et al., 2021).

2.5. Sequence similarity/identity and protein conservation across coronaviruses

This section discusses overall genomic conservation across the coronaviruses (CoVs). There is more than 70% genetic similarity between SC-2 and SC-1, which was responsible for the 2003 SARS outbreak. The pairwise amino acid sequence alignment pointed to sequence identity of 94.6% in seven conserved replicase domains of the ORF1ab polyprotein of SC-2 with SC-1 (Zhou et al., 2020). There is higher sequence homology between SC-2 and SC-1 than MERS-CoV (Xu et al., 2020). Studies have also suggested that SC-2 is closely related to bat SARS such as bat-SL-CoVZC45 (Lu et al., 2020) and bat-SL-CoVZXC21 (Chan et al., 2020). A study by Ceraolo and Giorgi (2020) found a 96.2% sequence identity of Bat-CoV RaTG13 genome (Gisaid EPI_ISL_402131) with the SC-2 reference genome (NC_045512.2) thus confirming the possibility of a zoonotic transmission.

The non-coding sequences, which are at the genomic terminals (5'-3' untranslated regions) of CoVs, have gene regulatory functions and are known to encode conserved RNA secondary structures (Chen and Olsthoorn, 2010). A study by Chan et al. (2020) used 109 representative CoV genomes collected from four genera (alpha, beta, gamma and delta) and the reference genome of SC-2 to find conserved features between them. MSA was performed using Clustal Omega (v1.2.4) and the results suggested that there was nucleotide sequence identity between SC-2 and the betaCoVs at various genomic locations such as 3'-UTR (97.4%), E gene (95.1%), ORF10 (93.8), 5'UTR (91.1%) and NSP10 (89.7%). The SC-2 3'UTR (approximately 30 nucleotides) also shared high sequence identities with deltaCoVs (97% from pigs) and gammaCoVs (94% from chicken and fowls). According to the study, the 3'UTR and 5'UTR are conserved within the betaCoV lineage B.

In another study the genome and protein sequences of various CoV isolates from humans, bats, civets and pangolin were downloaded from NCBI GenBank (Sayers et al., 2022)

(www.ncbi.nlm.nih.gov/sars-cov-2) as well as GISAID (Khare et al., 2021) (www.gisaid.org) database. The cDNA sequences from these databases were translated into protein sequences using the sequence analysis application MacVector. The sequences of protein were aligned using CLUSTAL W v.10, were compared to the reference isolate (Wuhan-Hu-1/2019) to identify mutations, indicating a 86% sequence identity of ORF1ab protein between SC-2 and SC-1 (Padhan et al., 2021).

2.6. Variants of SARS-CoV-2

The RNA viruses such as SC-2 are also known to adapt according to their new host which results in mutations and causes genetic evolution over time. These mutations eventually will branch out into different variants of the virus that are initially very similar to the ancestral strain but will tend to mutate further over time to produce more variable mutants. Many SC-2 variants have been reported since the first outbreak of the virus. Some of the variants impact on global health and are recognised as variants of concern (VOC) by the WHO. Regarding the different variants of SC-2, a study by Almubaid and Al-Mubaid (2021) analysed 1200 genomes sampled across the first seven months of 2020. They analysed mutations and mutation frequency/trend using sequence alignment tool Clustal Omega and Jalview for viewing the alignment. They reported that the ORF1a region mostly had synonymous mutations. Some synonymous mutations were also found in NSP12 and NSP13. Figure 2.3. shows MSAs of NSP1 of the SC-2 variants and the very high degree of conservation across this NSP.

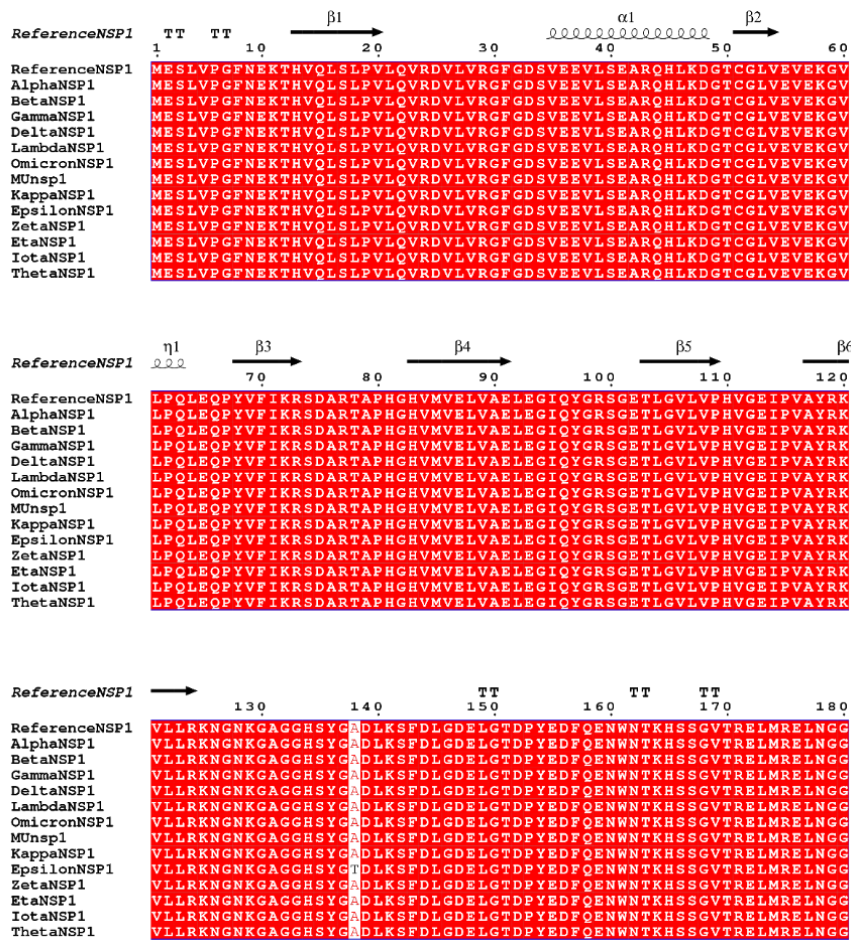


Figure 2.3. The MSA of NSP1 proteins with SC-2 Wuhan as the reference sequence and its variants (one sequence of each variant). The MSA of NSP1 was performed using Clustal Omega (Sievers et al., 2011) and rendered using ESPrpt version 3.0 (Robert and Gouet, 2014). Red colour shows sequence identity among the variants with the reference sequence. White colour shows mutation and deviation of amino acid sequence (presence of Threonine in Epsilon-NSP1 instead of Alanine).

Below are the major VOCs, that I considered in my work, however more sub-variants have emerged from these over time.

- Alpha (B.1.1.7): first variant of concern described in the United Kingdom (UK) in late December 2020
- Beta (B.1.351): first reported in South Africa in December 2020
- Gamma (P.1): first reported in Brazil in early January 2021
- Delta (B.1.617.2): first reported in India in December 2020
- Omicron (B.1.1.529): first reported in South Africa in November 2021

2.7. NSP conservation across coronaviruses (CoVs)

The **NSP1** of SC-2 has a sequence similarity of approximately 91% with NSP1 of SC-1. Considering structural similarity, the core domain of NSP1 is also similar between the two viruses (Clark et al., 2020; Semper et al., 2021). In a study, 47,427 NSP1 sequences of SC-2 were analysed against the Wuhan-Hu-1 reference strain. The sequence alignment was performed using the EMBOSS diffseq program which resulted in >97.6% sequence identity between the sequences (Min et al., 2020). NSP1 can halt host translation by interacting with the human 40S ribosomal subunit with the help of a dipeptide motif Lys164-His165 (Figure 2.4.) (Narayanan et al., 2008). These residues were found to be conserved in SC-1 and SC-2 in a sequence alignment of NSP1 C terminal (Schubert et al., 2020). The interaction between the C-terminal of NSP1 and the human ribosome should be further investigated to understand NSP1 mediated translation inhibition and development of inhibitors for same.

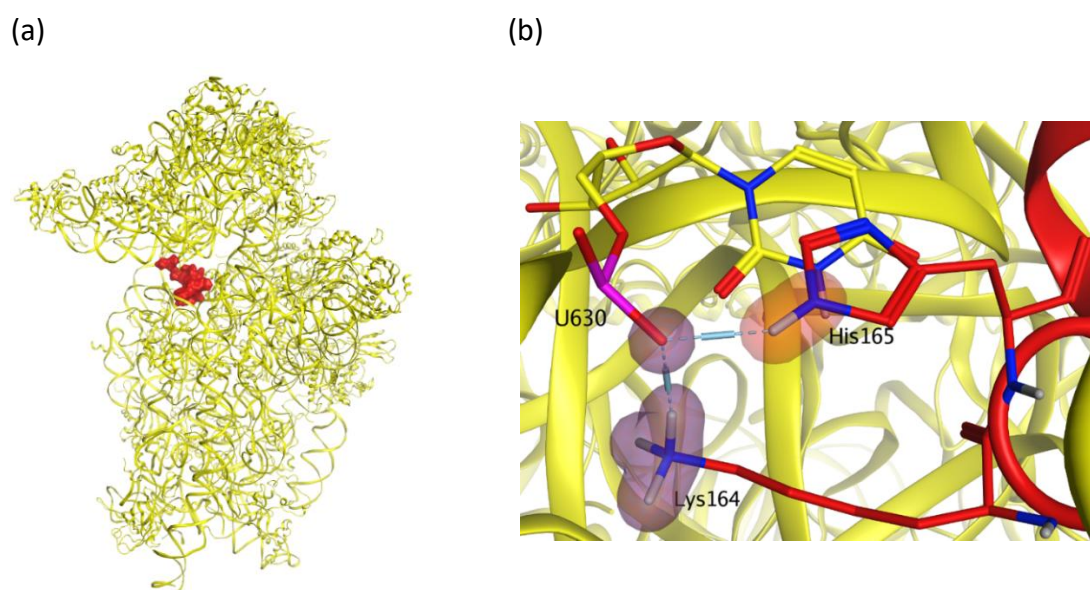


Figure 2.4. (a) C-terminal of NSP1 bound to the human 40S ribosomal subunit PDB ID- 6ZLW red: NSP1, yellow: ribosome (b) Shows the interactions of His165 and Lys164 with U630 of human ribosome. The picture was generated with MOE (2022).

The **NSP2** of SC-2 is less conserved than other NSP's, there is 68% identity with SC-1 NSP2 and just 20% identity with MERS-CoV NSP2. But there are four cysteine residues coordinating a Zn^{2+} ion in a Zn ribbon like motif which are conserved across these viruses (Verba et al., 2021).

NSP3 is the largest multi-domain protein produced by the coronaviruses. There is 83% sequence identity between the NSP3 of SC-2 and SC-1. In contrast, there is only 33% identity between NSP3 of SC-2 and MERS-CoV (Freitas et al., 2020). In a study the sequence alignment of the SC-1 and SC-2 SUD protein (one of the domains of NSP3) showed that the Lys565, Lys568 and Glu571 residues are conserved (Lavigne et al., 2021). X-ray structures of the NSP3 papain-like protease (PLP) domain, show that it has a homo-trimeric fold which is conserved with its SC-1 and MERS CoV counterparts. The key residues responsible for binding of ADP-ribose and other related adenosine-derivatives with NSP3-macrodomain X are conserved across beta-CoVs (Figure 2.5.). The structure also has high structural homology to the SC-1 NSP3-macrodomain X (Michalska et al., 2020).

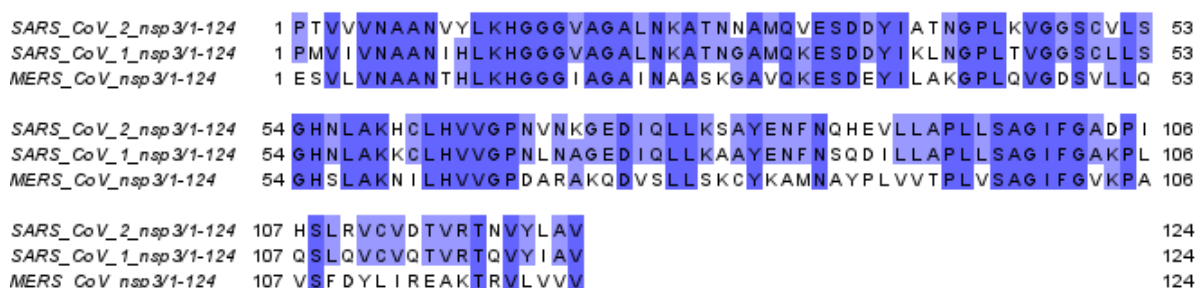


Figure 2.5. Sequence alignment using Clustal Omega and viewed with Jalview (Troshin et al., 2011, 2018) of NSP3 Mac1 from SC-2 (YP_009725299.1), SC-1 (NP_828862.2), and MERS (YP_009047231.1) showing percentage identity (purple colour to white as identity decreases).

NSP4, a transmembrane protein, is a component of the replication complex found in CoVs and it has a sequence identity of 80% between different SARS strains (Davies et al., 2020). A study by Gordon et al. (2020) reported 90.8% sequence similarity between NSP4 of SC-2 and SC-1. The C-terminal of NSP4 is found to be conserved among *Nidovirales* (Neuman, 2016).

The viral protease **NSP5**, is highly conserved, has a chymotrypsin-like fold (3CLpro) and is also known as the main protease (Mpro) (Anand et al., 2002, 2003; Gorbalenya et al., 1989). There is high tertiary and quaternary structural conservation especially in domains 1 and 2 of Mpro across coronaviruses. The N-terminal domains (1 and 2) forms a

chymotrypsin-like fold consisting of beta-barrels which is known to be highly conserved among coronaviruses (Stobart et al., 2013). Two highly conserved residues (His41Cys145) are present at the catalytic site of the SC-2 and SC-1 protease (Shitrit et al., 2020). A review by Goyal and Goyal (2020) discussed the importance of targeting the dimerization of the main protease as a broad-spectrum therapeutic strategy. The residues involved for dimerization are highly conserved across beta coronaviruses (Figure 2.6.).

(a)

SARS_CoV_2_nsp5/1-306	1	SGFRKMAFP	SGKVEG	CMVQVTCGTTTLN	GLWLDD	VVYCPRHVICT	SEDMLNPN	53		
SARS_CoV_1_nsp5/1-306	1	SGFRKMAFP	SGKVEG	CMVQVTCGTTTLN	GLWLDD	TVYCPRHVICT	TAEDMLNPN	53		
MERS_CoV_nsp5/1-306	1	SGLVKMSHP	SGDVEA	CMVQVTCGSMITL	NGLWLD	NTVWCPRHVMCP	ADQLSDPN	53		
SARS_CoV_2_nsp5/1-306	54	YEDLLIRKSNHN	FLVQA	GNVQLRVIGHSMQNC	VLLK	KVD	TANPKTPKYKF	103		
SARS_CoV_1_nsp5/1-306	54	YEDLLIRKSNH	SFLVQA	GNVQLRVIGHSMQNC	LLK	KVD	TANPKTPKYKF	103		
MERS_CoV_nsp5/1-306	54	YDALLISMTNHS	FSVQKH	IGAPANLRV	VGHAMQ	STLLK	LTVDVANPSTRAYTF	106		
SARS_CoV_2_nsp5/1-306	104	VRIQPGQTF	SVLACYNGSP	SGVYQCAMPNF	TIKGS	FSLNGSCGS	SVGFNIDYDC	156		
SARS_CoV_1_nsp5/1-306	104	VRIQPGQTF	SVLACYNGSP	SGVYQCAMPNHT	IKGS	FSLNGSCGS	SVGFNIDYDC	156		
MERS_CoV_nsp5/1-306	107	TTVKPGAAF	SVLACYNGR	PTGTF	TVVMRPNY	TIKGS	FSLNGSCGS	SVGYTKEGSV	159	
SARS_CoV_2_nsp5/1-306	157	VSFCYMHMEL	PTGVHAGTDLEG	NFYGPF	VDRQTAQAAG	TDTTIT	TVNVLAWLY	209		
SARS_CoV_1_nsp5/1-306	157	VSFCYMHMEL	PTGVHAGTDLEG	KFYGPF	VDRQTAQAAG	TDTTIT	TVNVLAWLY	209		
MERS_CoV_nsp5/1-306	160	INFCYMHQMEL	ANGTHTG	SAFDG	TMYGAFMDKQ	VHQVQL	TDKYCSVNVV	AWLY	212	
SARS_CoV_2_nsp5/1-306	210	AAVINGDRWF	LNRFTTTLNDF	NLVAMKYNYEPL	TQDHVD	ILGPL	SAQTGIAVL	262		
SARS_CoV_1_nsp5/1-306	210	AAVINGDRWF	LNRFTTTLNDF	NLVAMKYNYEPL	TQDHVD	ILGPL	SAQTGIAVL	262		
MERS_CoV_nsp5/1-306	213	AAILNGCAWF	VKPNRTSVVS	FNEWALANQF	TEFVG	TQ--	SVDM	LAVK	TGV	263
SARS_CoV_2_nsp5/1-306	263	DMCASL	KLKELLQNGMNGRT	ILGS	ALLEDE	FTFPD	VVRQCSGVTFQ	306		
SARS_CoV_1_nsp5/1-306	263	DMCAAL	KLKELLQNGMNGRT	ILGST	ILEDE	FTFPD	VVRQCSGVTFQ	306		
MERS_CoV_nsp5/1-306	264	QLLYAI	QQLYTQFGQKQ	ILGST	MLED	FTFPD	VNMQIGVVMQ	306		

(b)

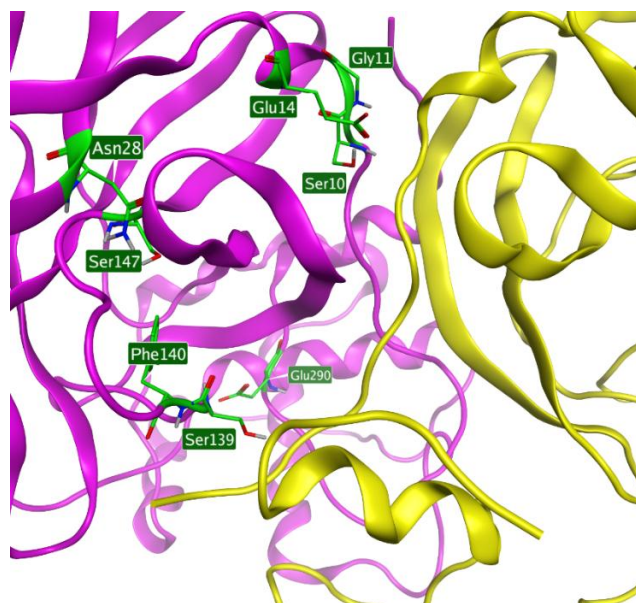


Figure 2.6. (a) Sequence alignment using Clustal Omega and viewed with Jalview of NSP5 from SC-2 (YP_009742612.1), SC-1 (YP_009944370.1), and MERS (YP_009047233.1) showing percentage identity (purple colour to white as identify decreases). Red boxes highlight the conserved residues involved in Mpro dimerization. (b) SC-2 Mpro (PDB ID:

1UK4) showing the two monomers A and B of the dimer in pink and yellow colour respectively. The conserved residues of monomer A are highlighted in green colour (Ser10, Gly11, Glu14, Asn28, Ser139, Ser144, Ser147, Glu166, Glu290 and Arg298) that can be targeted to inhibit the dimerization of SC-2 Mpro. The picture was generated with MOE (2022).

A recent study found regions conserved within **NSP6** of the Coronaviridae family which could be targeted for developing a new therapeutic approach (Gupta et al., 2020). These conserved regions are mostly coiled secondary structure with multiple charged and aromatic amino acids which are responsible for protein-protein interactions. The MSA of 10,664 SC-2 virus genomes across 73 countries confirmed the global stability of conserved regions within ORF1ab specifically in the NSP6 region (Saha et al., 2021).

There were highly conserved residues found to be located at the two oligomer interfaces of **NSP7** and NSP8 (Biswal et al., 2021). The replication-transcription complex of NSP12-7-8 have high conservation of the structural features between SC-1 and SC-2 (Kirchdoerfer and Ward, 2019; Gao et al., 2020)

A structure-based sequence analysis study suggested that the residues of **NSP8** forming interactions with NSP7 are highly conserved among coronaviruses (Biswal et al., 2021). A study by Gordon et al. (2020) reported to have 99% sequence similarity between NSP8 of SC-2 and SC-1.

For the replication process, the **NSP9** monomer of CoVs is also very important. An MSA study of NSP9 found that there are two glycine residues (G100 and G104) which are conserved across CoVs (Miknis et al., 2009). A study by Littler et al. (2020) suggested that there is a high level of structural and sequence (97%) conservation of NSP9 between SC-2 and SC-1. The study also reported that NSP9 has a conserved GxxxG interaction motif (the same glycine residues highlighted above) within the α -helix which facilitates a mini coiled-coil dimerization interaction.

A study revealed that **NSP10** acts as a co-factor to NSP14 and NSP16 in SC-1 thereby regulating their ExoN and ribose-2'-O-MTase (2'-O-MTase) activities (Bouvet et al.,

2010, 2012). NSP10 is highly conserved among lineage B of the beta-coronaviruses. A study by Rosas-Lemus et al. (2020) found that after performing sequence alignment of NSP10 from Bat-CoV-RaTG13, SC-1, Bat-SARS-like coronavirus (Bat-SL-CoV) isolate Rs4247 and MERS-CoV, the sequence identity of SC-2 NSP10 with these isolates were 100%, 99%, 98% and 59% respectively. After analysing the aligned sequences, it was found that zinc coordinating residues C74, C77, H83, C90, C117, C120, C128 and C130 were 100% conserved, thus indicating to the importance of zinc binding activity of NSP10 which is required for viral replication-transcription processes.

The SC-2 viral polymerase RdRp (**NSP12**) helps in replicating the viral genome. Conserved replicase domains indicate common ancestry of the *nidoviruses* (Gorbalenya et al., 2006). NSP12 also shares conserved active residues with other human coronaviruses such as SC-1, MERS-CoV, HCoV-OC43 and HCoV-229E (Dhankhar et al., 2021) (Figure 2.7.). There is the presence of seven conserved polymerase motifs (A to G) in the catalytic core of NSP12 of SC-2 residing in the palm domain resembling NSP12 of SC-1 (Kirchdoerfer and Ward, 2019).

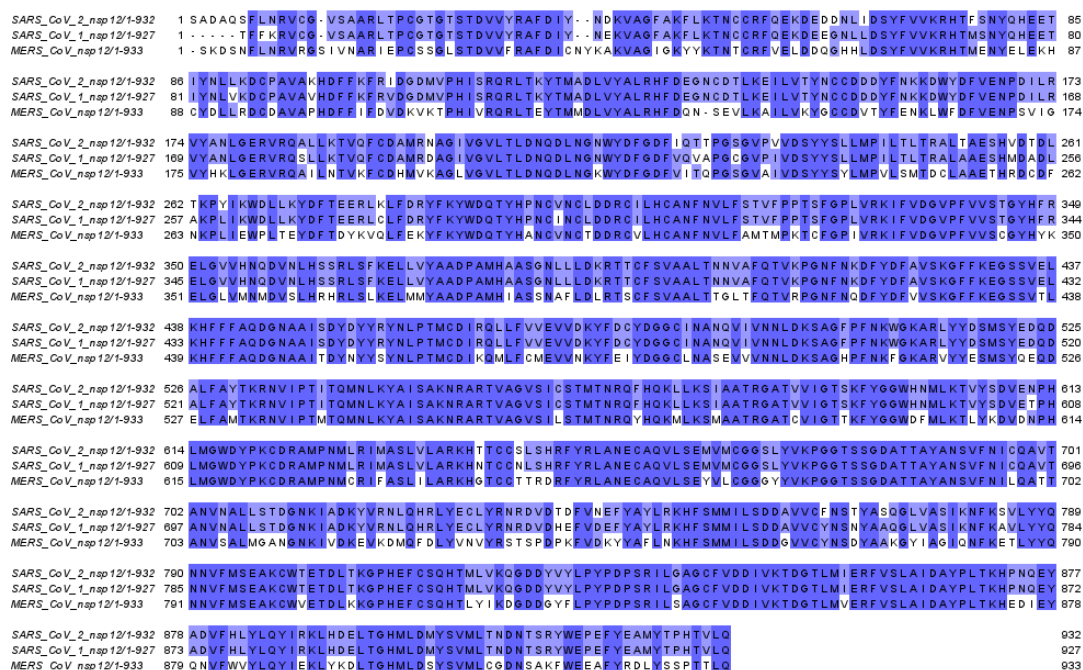


Figure 2.7. Sequence alignment using Clustal Omega and viewed with Jalview of NSP12 from SC-2 (YP_009725307.1), SC-1 (NP_828849.7), and MERS (YP_009047223.1) showing percentage identity (purple colour to white as identity decreases).

An interesting study by Yazdani et al. (2021), performed sequence alignments of 15 SC-2 proteins and compared the variations across 27 α - and β -coronaviruses. They also mapped

druggable binding pockets on the protein structures. The results suggested that a druggable channel overlapping with RNA binding site of helicase protein (**NSP13**) of SC-2 was highly conserved i.e., >90% of the residues lining the pocket were conserved across 27 α - and β -coronaviruses. The helicase protein is highly conserved across SC-2, SC-1 and MERS-CoV (Figure 2.8.). This channel has recently been chosen by the CACHE consortium as the target for its second international hit finding competition (<https://cache-challenge.org/>). Another study found that the residues in the zinc-binding domain of NSP13 making interactions with NSP8 of the replication complex are highly conserved in *Betacoronaviruses* (Chen et al., 2020).

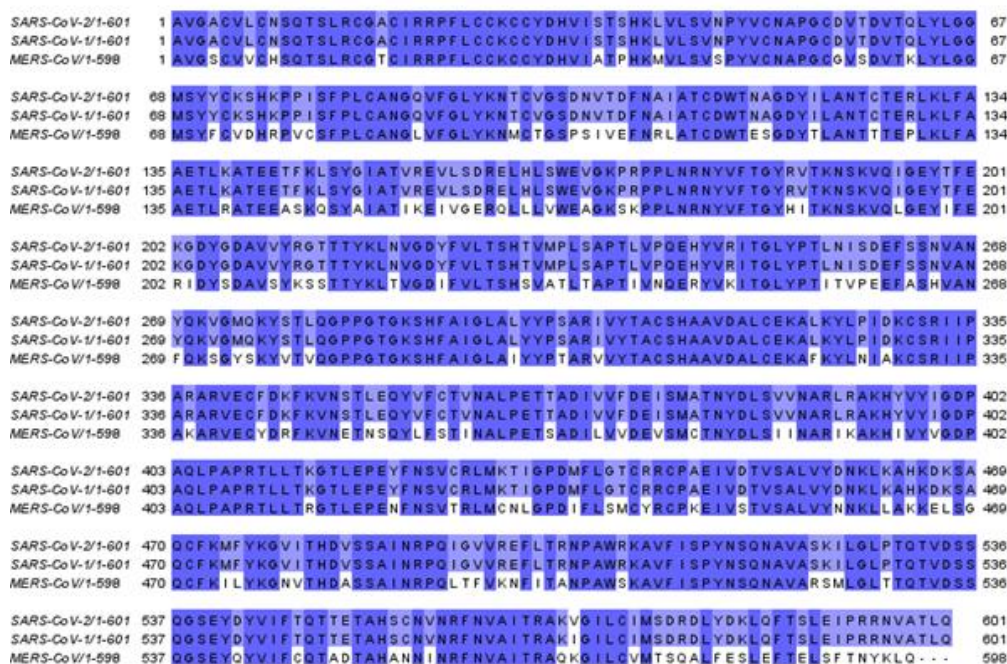


Figure 2.8. Sequence alignment using Clustal Omega and viewed with Jalview of NSP13 from SC-2 (YP_009725308), SC-1 (NP_828870.1), and MERS (YP_009047224.1) showing percentage identity (purple colour to white as identity decreases).

NSP14 acts as an exoribonuclease and N7-MTase in CoVs. An MSA (using CLUSTALW) study of NSP14 protein sequences was performed among different genera: SC-1 (genus *Betacoronavirus*), HCoV-229E (genus *Alphacoronavirus*) and IBV (genus *Gammacoronavirus*) which indicated the presence of conserved residues (Ma et al., 2015). The presence of a conserved DEDDh motif in the active site of NSP14 is known to be distributed over three canonical motifs (I, II and III). The SC-2 residues D90/E92 (motif I), D243 (motif II), and D273 (motif III) were reported to be fully conserved (Figure 2.9.) in the other CoVs (SC-1 and MERS-CoV) (Saramago et al., 2021).

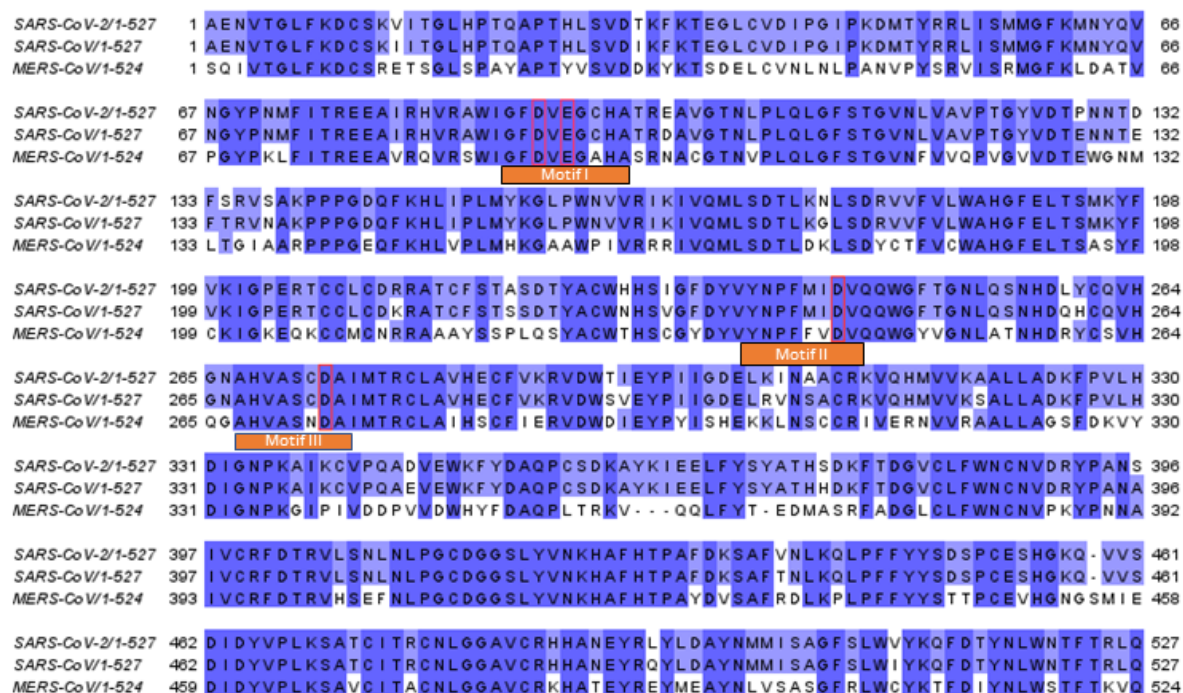
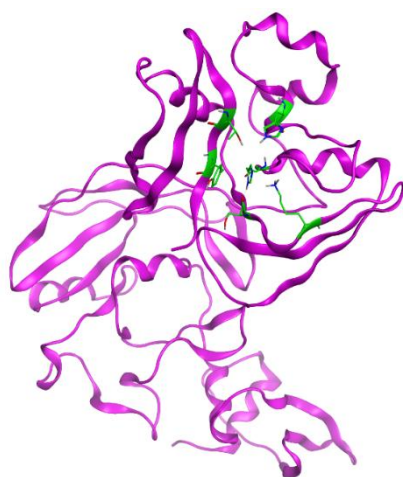


Figure 2.9. Sequence alignment Using Clustal Omega and viewed with Jalview of NSP14 from SC-2 (UniProt ID: P0DTD1), SC-1 (UniProt ID: P0C6X7) and MERS (YP_009047225) showing conserved residues (red box) within the different motifs, colour scheme (purple) = percentage identity.

The **NSP15** of coronaviruses is known to have endoribonuclease activity and to be conserved in coronaviruses (Snijder et al., 2003). The residues H235, H250, K290 and T341 present in the catalytic site of the NendoU domain are known to be highly conserved in SARS-CoVs (Joseph et al., 2007). There is 94.8% sequence similarity between the NSP15 of SC-2 and SC-1 (Saramago et al., 2022). The six key residues His235, His250, Lys290, Thr341, Tyr343 and Ser294 (Figure 2.10.) which are present in the active site of NSP15 are conserved among SC-2, SC-1 and MERS-CoV (Kim et al., 2020).

(a)



(b)

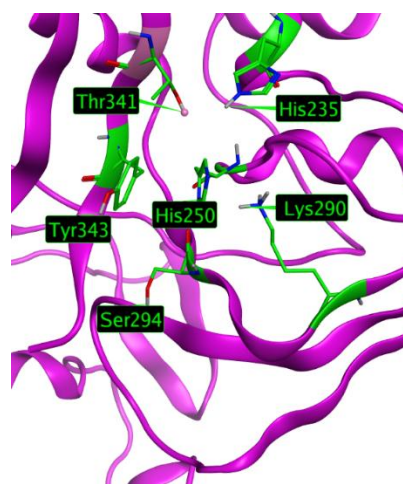


Figure 2.10. The conserved residues (green colour) present inside the active site of NSP15 (pink colour) of SC-2 (PDB ID: 6WLC). The picture was generated with MOE (2022).

The structure and function of **NSP16** is highly conserved across CoVs. The sequence alignment study by Rosas-Lemus et al. (2020) of SC-2 NSP16 with Bat-CoV-RaTG13, Bat-SL-CoV Rs4247, SC-1 and MERS-CoV resulted in sequence identity percentages of 100%, 93%, 95% and 66% respectively. Some highly conserved amino acid residues were Lys46, Asp130, Lys170 and Glu203 which are present in the SAM Mtase binding sites of NSP16 (Figure 2.11.).

(a)



(b)

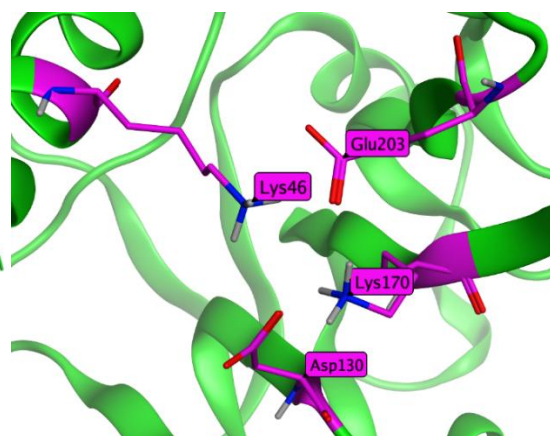


Figure 2.11. The conserved residues (pink colour) present in the Mtase binding site of NSP16 (green colour) of SC-2 (PDB ID: 6YZ1). The picture was generated with MOE (2022).

2.8. Material and methods

2.8.1. Bioinformatics materials

- **Databases:** National Center for Biotechnology Information (NCBI-www.ncbi.nlm.nih.gov) and Global Initiative for Sharing All Influenza Data (GISAID-<https://gisaid.org>)
- **Software:** Kalign 3.3.2, Mafft version 7.503, Clustal Omega version 1.2.4, T-coffee version 13.45.0, Muscle version 3.8.1551, Jalview version 2.11.26, Ugene version 41.0, MEGA11, Ubuntu version 20.04.6, BioAider version 1.423, Notepad++ version 8.4.6, Microsoft excel, Windows Subsystem for Linux (WSL), python version 3.8.10, MobaXterm version 21.5, Sublime text editor version 4143, Molecular Operating Environment (MOE) version 2022.02, PyMOL 2.5.8.
- **Python package:** Seqtk
- **Scripts:** Python script for MSA analysis, Python script for MSA chopping script, Python script for count plot
- **Operating systems:** Microsoft Windows 11 Education, Linux OS
- **System:** Dell Precision 3650 Tower, Processor- 11th Gen Intel(R) Core™ i9-11900 @ 2.50GHz, 2496 Mhz, 8 Core(s), 16 Logical Processors with NVIDIA graphic card
Servers: Stibbons - AMD Opteron (TM) processor 6276 with 64 cores running on Ubuntu 20.04.6, Irish Centre for High-End Computing (ICHEC supercomputer)

2.8.2. Benchmarking of MSA tools methods

Before getting started with the bioinformatics analysis and performing the alignments, benchmarking of different MSA tools was performed. Kalign, Mafft and Clustal Omega was used for aligning different datasets (200, 1000, 2000, 3000, 5000, 10,000, 50,0000, and 92,989) of sequences. The sequences were downloaded from NCBI with the following filters: protein- ORF1ab, length- 7096, host- human, X characters- 0 and search for complete sequences. For all the alignments the calculations were run over 14 cores. All the MSA tools were installed on Linux operating system and were accessed using WSL or Ubuntu. Other MSA tools such as T-coffee and Muscle that work on progressive and iterative progressive alignments respectively were also used to align 200 sequences. The

time taken by each tool to align sequences were recorded and graphs were plotted using MS Excel.

Kalign command: *kalign -i input.fasta -o output.fasta -nthreads 14*

Clustal Omega command: *clustalo -i input.fasta -o output.fasta -threads=14*

Mafft command: *mafft --clustalout --thread 14 input.fasta > output.fasta*

Muscle command: *muscle -align input.fasta -output output.fasta*

T-coffee command: *t_coffee input.fasta*

2.8.3. NCBI and GISAID databases for downloading sequences

The SC-2 ORF1ab amino acid sequences were first downloaded from NCBI SARS-CoV-2 resources database. The date of download was on the 21st of February 2022 with relevant filters such as nucleotide completeness – complete, protein - ORF1ab, host – humans. It resulted in **1,070,921** sequences. Two types of files were downloaded; (1) FASTA file having all the sequences; and (2) CSV file. **Note: This dataset was used for statistical and conservation analysis of all the NSPs.**

The allprot0730.tar.xz file was downloaded from GISAID on the 01st of August 2022. The allprot.tar file had **11,910,751** amino acid sequences of the entire SC-2 genome. The dataset downloaded had different headers which were difficult to process during the statistical analysis. **Note: This dataset was only used for conservation analysis of NSP3 and NSP13 (Discussed in Chapter III).**

2.8.4. Statistical analysis on NCBI dataset

For performing the statistical analysis, the CSV file downloaded from NCBI was used. Separate text files were then generated having only one column i.e., Geo location, length and Pangolin. Standard Linux commands were used for counting and sorting the entries. The final output files were then converted into CSV files. The graphs were generated using python library such as matplotlib.

Linux commands:

```
>> sort <filename> |
```

```
>> uniq -c > output1.txt (for counting)
```

```
>> sort -r <output1.txt> > output2.txt (for arranging in descending order)
```

Python commands:

```
>>> import matplotlib.pyplot as plt
```

```
>>> import pandas as pd
```

```
>>> data = pd.read_csv(r"D:\Documents\input.csv")
```

```
>>> df = pd.DataFrame(data)
```

```
>>> df
```

```
>>> X = list(df.iloc[:,0])
```

```
>>> Y = list(df.iloc[:,1])
```

```
>>> plt.bar(X, Y, color='g')
```

```
>>> plt.title("XYZ vs Count")
```

```
>>> plt.xlabel("XYZ")
```

```
>>> plt.ylabel("Count")
```

```
>>> plt.xticks(rotation='vertical')
```

```
>>> plt.rcParams["figure.figsize"] = (15,10)
```

```
>>> plt.show()
```

2.8.5. Phylogenetic tree on NCBI dataset

A dataset was prepared having SC-2 ORF1ab sequences from different variants and phylogenetic tree was generated using MEGA11. The build tree option was selected from the main menu. The phylogenetic tree was generated based on Maximum Likelihood tree building method and the distance matrix model used was Whelan and Goldman (WAG) with gamma distribution (5 discrete categories). The bootstrap method was used for running 200 replicates.

2.8.6. Pre-processing of NCBI dataset

Python script (msa_result_condence.py by Dr. Karsten Hokamp https://github.com/khokamp/python-scripts/blob/main/msa_result_condense.py), bash commands and BioAider outputs were used for removing sequences having BJXZ ambiguous characters and uncommon amino acids such as OU. Highly dissimilar sequences that generated large gaps within the alignment were also removed. The unknown amino acid characters (BJOUXZ) were identified by first replacing them with a question mark (?) and then removing the sequences having that question mark using bash commands.

Bash commands: Removed sequences containing BJOUXZ characters.

1. *perl -pe 's/[BJOUXZ]/?/gi unless m/>/;' input.fasta > output.fasta*
2. *\$cat input.fasta |*
> sed -e 's/>/>/g' -e 's/\$/#/g' |*
> tr -d '\n' |
> tr '' '\n' |*
> fgrep -v ? |
> tr '#' '\n' |
> sed 1d > output.fasta

Sequences with an extra amino acid or a single different amino acid were removed, as they were creating large gaps (outlier sequences) that made the MSA messy and difficult to analyse. These sequences were removed by first performing MSA with these being present within the dataset. These sequences were then identified using the output files generated by the MSA analysis Python script and BioAider log file, indicating the residue position along with the sequences associated (accession number). These dissimilar sequences were then removed from the datasets. After the preprocessing step there were **1,070,751** sequences.

2.8.7. Pre-processing of GISAID dataset and statistical analysis

The GISAID dataset had all the SC-2 genome sequences within the allprot.txt file. Linux commands were used for first extracting the sequence headers followed by the extraction of locations (countries) from the header pipe. The output file was then used for counting

and sorting the entries. The final output file was converted into CSV file. The location vs count plot was generated using the python matplotlib library as mentioned in **2.8.4**.

```
>> grep -e ">" nsp1_prot.fasta > only_header_for_stats_nsp1.txt (To extract only the headers).
```

```
>>> awk -F"/" '{print $0">"$pipe_location_add_number}' input.txt (To capture particular pipe)
```

```
>> sort <filename> |
```

```
>> uniq -c > output1.txt (for counting)
```

```
>> sort -r <output1.txt> > output2.txt (for arranging in descending order)
```

Pre-processing of dataset:

The GISAID dataset has sequences with different AA lengths. A python bioconda package called seqtk (<https://github.com/lh3/seqtk>) was used to extract sequences having a certain range of length for each NSP. The range was selected considering the actual AA length of the NSP.

```
>>> seqtk seq in.fasta -m minimum length -M maximum length > out.fasta
```

The sequences submitted on GISAID have headers that extend onto the next line which might interfere with the sequence reading and result in MSA error. Hence, a bash command was used to extract the accession number of sequences from the header pipe.

```
>>> awk -F"/" '{print $0">"$pipe_location_add_number}' input.txt (To capture particular pipe)
```

The next step was then to remove sequences having unknown amino acid characters. Bash commands were used to remove BJOUXZ characters as mentioned in **2.8.6**.

A few of the steps can create empty lines. The Sublime text editor was used for removing those empty lines.

- Go to Find → Replace
 - or – press the Ctrl + H on Windows/Linux or  Cmd + Alt + F on Mac
- Select the 'Regular expression' by clicking on the .* button in the Find box
 - or – press the Alt + R on Windows/Linux or  Cmd + Alt + R on Mac

- In the Find field, type in:
 $\text{^\n}$ to delete empty lines excluding the ones that contain whitespaces:
 – or – $\text{^\(r|\n\r?)}$ to delete empty lines including ones with whitespaces:
- Leave the Replace field empty
- Click on the Replace All button

For removing highly dissimilar sequences the workflow in **2.8.6.** was followed.

2.8.8. Variant distribution

The NCBI dataset was used to extract the Pangolin lineages from the sequence headers using bash commands (**Appendix 6.2.**). The header information was then matched with the variants of SC-2 (Alpha, Beta, Gamma, Delta, Omicron and Non-VOCs). The CSV of the variants was obtained from <https://3dgiordano.github.io/SARS-CoV-2-Variants/>. The variant distribution bar graph was generated using MS Excel.

2.8.9. MSA using Kalign and visualisation of conserved/mutated amino acids

The MSAs were conducted on both the datasets using Kalign with the command as mentioned in **section 2.8.2.** and the sequence analysis was carried using the `msa_result_condence.py` Python script by Dr. Karsten Hokamp. The reference sequence used for the MSA was YP_009724389 (Wuhan reference) and was placed at the top of the input FASTA file. As the MSA was performed on the entire polyprotein i.e., ORF1ab it was necessary to chop the entire MSA into individual 1-16 nsp MSA's for the identification of conserved/mutated regions among them. The python script (`msa_result_condence.py` by Dr. Karsten Hokamp) reads in the reference FASTA file (Wuhan reference) having sequences of each nsp's and the MSA file which results in the chopping of the entire MSA into individual nsp MSA files. After the completion of the MSA the output was analysed using a python script.

```
>>> python3 msa_result_condence.py msa_output_file.fasta nsp1_16_reference.fasta
```

An excel sheet with details about each AA residue position (gaps, percent of conservation, mutations) is also generated after chopping.

- **Analysis**

MS Excel was used to perform analysis on the python script output excel file. The conditional formatting with three colour scale was used for highlighting highly conserved (green), intermediate conserved (yellow) and highly mutated (red) AA residues. The conservational and mutational threshold values were selected arbitrarily for each NSP. Different clusters were created according to these thresholds, i.e., conserved cluster, intermediate cluster and mutated cluster.

BioAider (Zhou et al., 2020) was also used to perform mutational analysis to identify each type of AA mutation along with a log file with sequence accession number having that mutation.

Note: The MSA file should have the reference (Wuhan) sequence at the top.

- **Visualisation of conserved amino acids**

Molecular Operating Environment (MOE) version 2022.02 (MOE, 2022) was used to visualise the conserved and mutated amino acid residues. The protein structure was prepared using the MOE quick prep option with the default Amber10:EHT force field, which considers explicit hydrogen atoms, tautomeric states and possible breaks in protein structure prior to conducting restrained all-atom molecular mechanics minimisation alongside electrostatics calculations. The sequence alignment tool was used to select the amino acids and label/colour them using the residue IDs. For incomplete structures loop (100 outputs) modelling was performed using MOE by selecting the unresolved amino acid residues. The top ranked modelled structure was selected for the 3D visualisation. There were a few NSPs which did not have a resolved structure and in that case AlphaFold (Jumper et al., 2021) was used. The surface was generated using the protein residues for visualisation. An *in-house* python script was also utilised to automatically colour the highly conserved amino acid as green, mutated as red and intermediate conserved as white on PyMOL. The site finder tool of MOE was also used to identify putative binding pockets. The tool assigns hydrophilic and hydrophobic cavity point also known as dummy atoms to possible binding pockets.

2.9. Results

2.9.1. MSA tools benchmarking

After performing the benchmarking of different MSA tools i.e., Kalign, Mafft and Clustal Omega (Figure. 2.12.) it was found that Kalign was the fastest among them in aligning sequences. Mafft and Kalign took almost same time in aligning 200-50,000 sequences but for aligning 92,989 sequences Kalign (took 1 hour 40 mins) was found to be faster than Mafft (took 3 days).

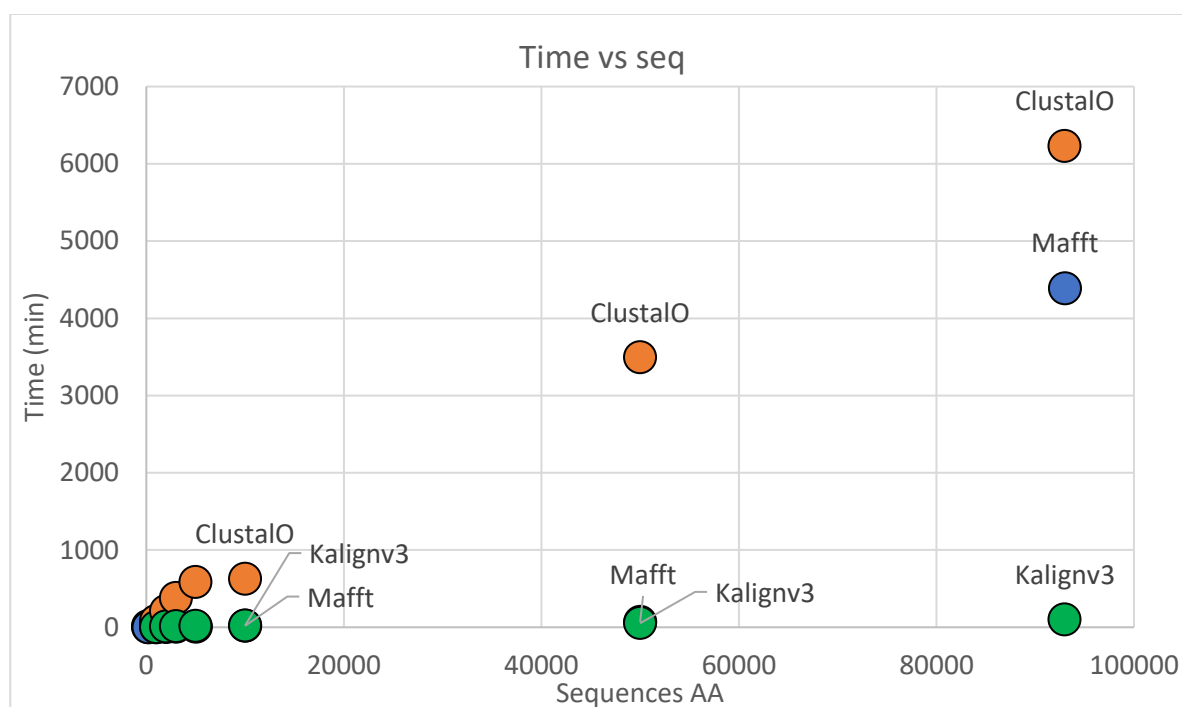


Figure 2.12. Graph showing time (min) on y-axis vs number of sequences aligned on x-axis by different MSA tools. Orange- Clustal Omega, Blue- Mafft and Green- Kalign.

The benchmarking of Kalign, Mafft, Muscle and T-coffee MSA tools on 200 sequences (Figure 2.13.) also suggested that Kalign (1 min) and Mafft (seconds) were the fastest in aligning sequences. This benchmarking was performed to see the time difference between MSA tools and their different alignment algorithms impacts the MSA completion time.

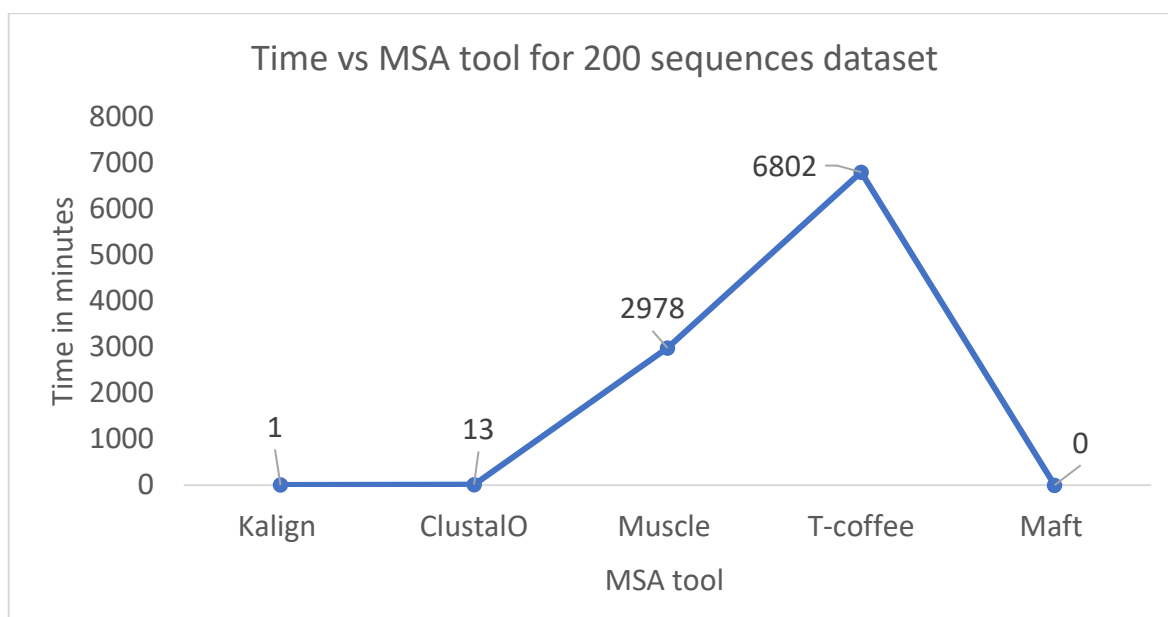
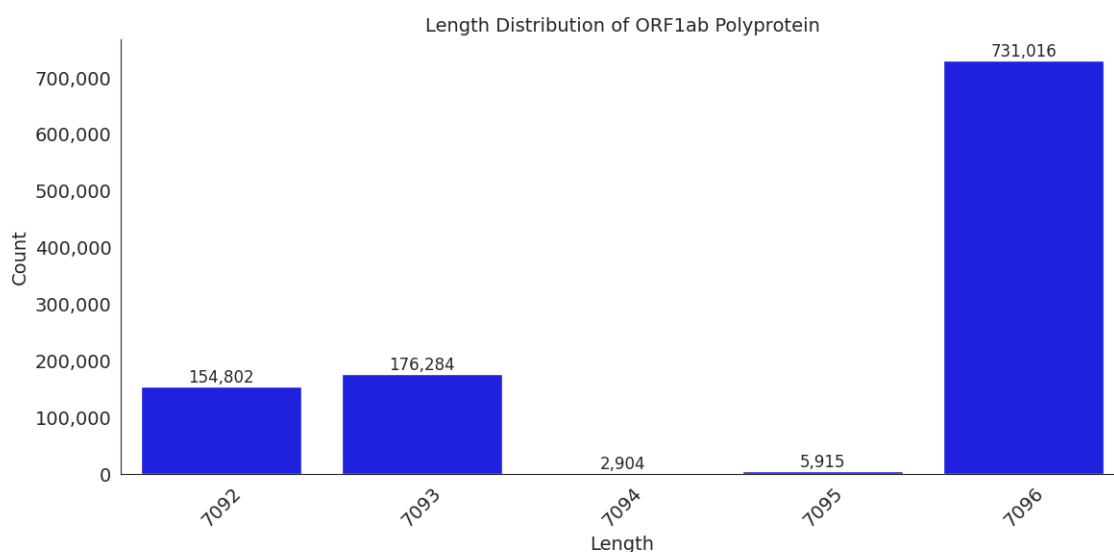


Figure 2.13. Time taken by different MSA tools to align 200 protein sequences. The x-axis represents the different MSA tools, and the y-axis represents time (mins) taken to align the 200 sequences.

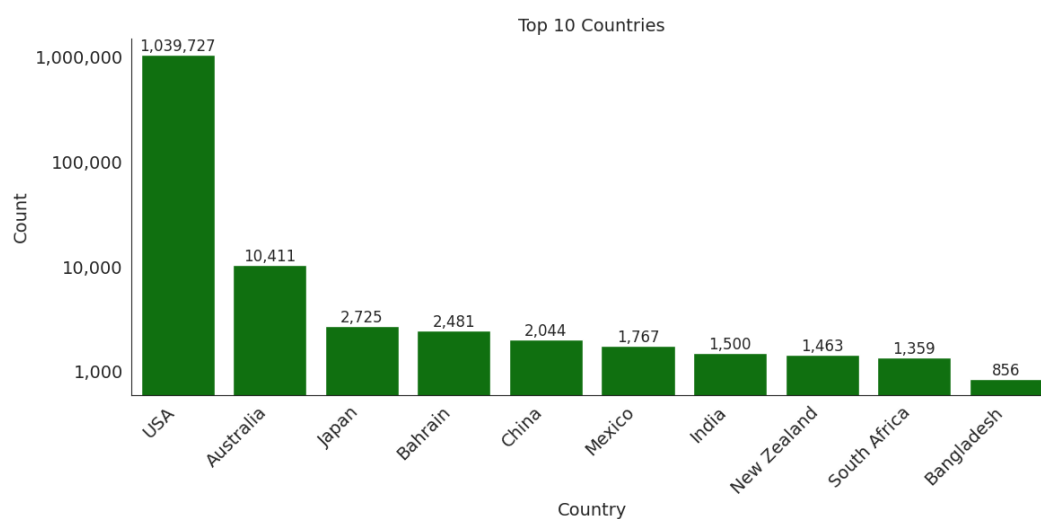
2.9.2. Statistical analysis on NCBI dataset

The dataset from NCBI had major submissions of ORF1ab sequences having length of 7092, 7093 and 7096 amino acids (Figure 2.14. a). Among the countries, USA had the most sequence submissions followed by Australia (Figure 2.14. b). Most of the sequences belonged to AY.103 Pangolin lineage (Figure 2.14. c).

(a)



(b)



(c)

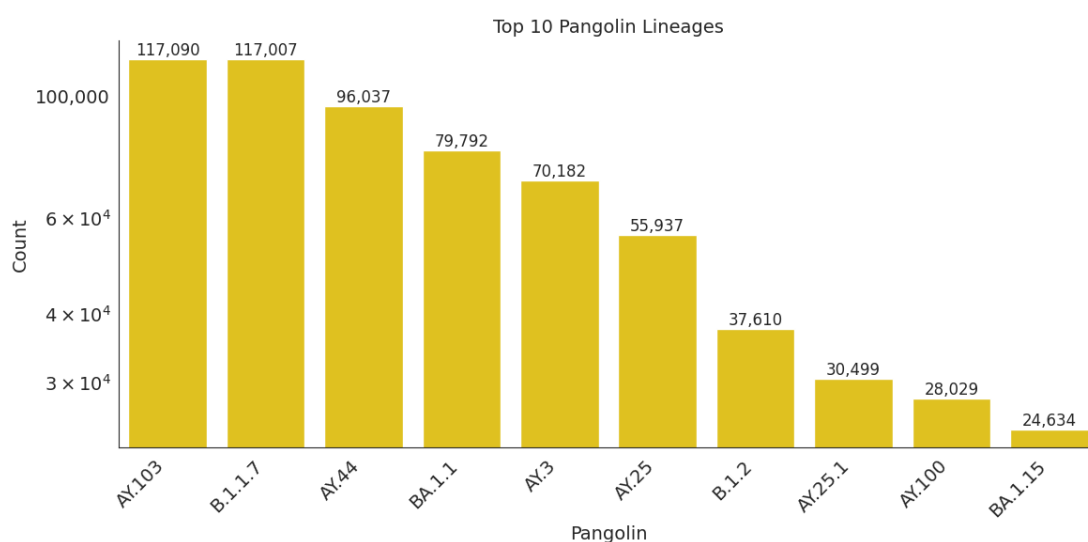


Figure 2.14. (a) Count plot of ORF1ab AA sequence length where y-axis shows the number of sequences, and x-axis shows length of AA sequence. (b) Count plot of geographical location where y-axis shows the number of sequences, and x-axis shows the geographical location. (c) Count plot of SC-2 variants where y-axis shows the number of sequences, and x-axis shows the pangolin lineages.

2.9.3. Phylogenetic tree of NCBI dataset

The phylogenetic tree indicated that most of the variants originated from the Wuhan variant (reference) (Figure 2.15.). The bootstrap values at each clade indicated high confidence in the placements of these variants except for the ones that had 24, 25, 28, and

30 values. The BA.2 and B.1.1.529 cluster near each other, and they share a more recent ancestor. The Epsilon variants are close together and likely evolved from a common ancestor. The Delta variant B.1.617.2 forms its own clade as it significantly diverged from earlier variants but still maintains some degree of similarity with Omicron subvariants. As this phylogenetic tree is an amino acid-based evolutionary distance tree rather than a time-scale tree, the placement of the variants is not in the chronological order of emergence.

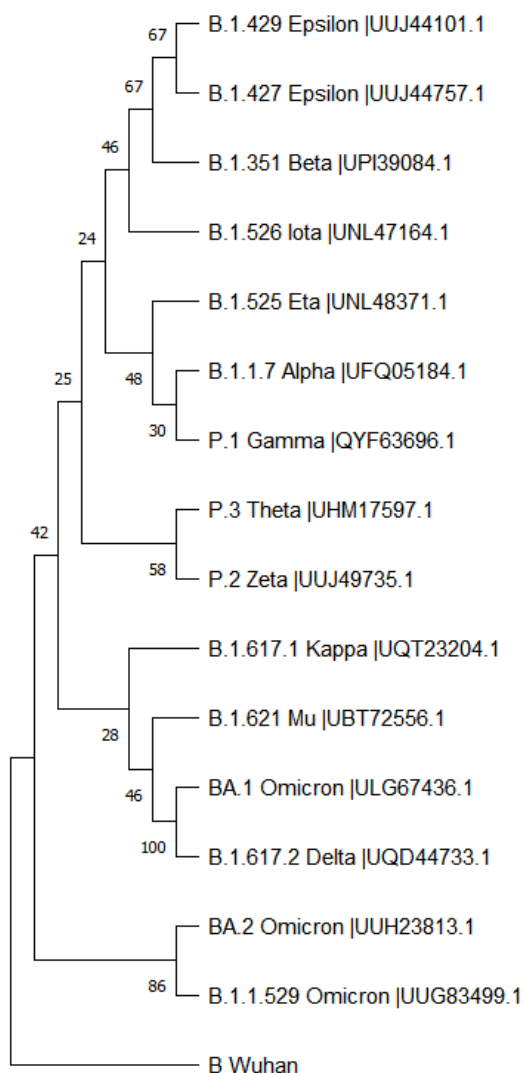


Figure 2.15. The evolutionary relationships among representative SC-2 lineages are illustrated based on amino acid sequence alignment. The tree includes the original Wuhan reference strain (B. Wuhan) and variants of concern, such as Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), and Omicron (B.1.1.529, BA.1, BA.2). It also encompasses other variants of interest, including Epsilon (B.1.427/B.1.429), Iota (B.1.526), Eta (B.1.525), Theta (P.3), Zeta (P.2), Kappa (B.1.617.1), and Mu (B.1.621). The

bootstrap values are provided at the branch nodes to indicate the confidence level in the branching order. GenBank accession numbers for each sequence are listed alongside their respective lineage names.

2.9.4. Statistics on GISAID dataset

The dataset from GISAID had most submissions from the USA, followed by the United Kingdom (Figure 2.16.). This made the dataset more diverse than the NCBI dataset, considering the more widely spread sequence submissions from around the world. Even though the most sequence submissions were still made by the USA, there is a substantial number of submissions made by other countries, which helps in avoiding bias in the data due to SC-2 variants circulating in specific geographical areas.

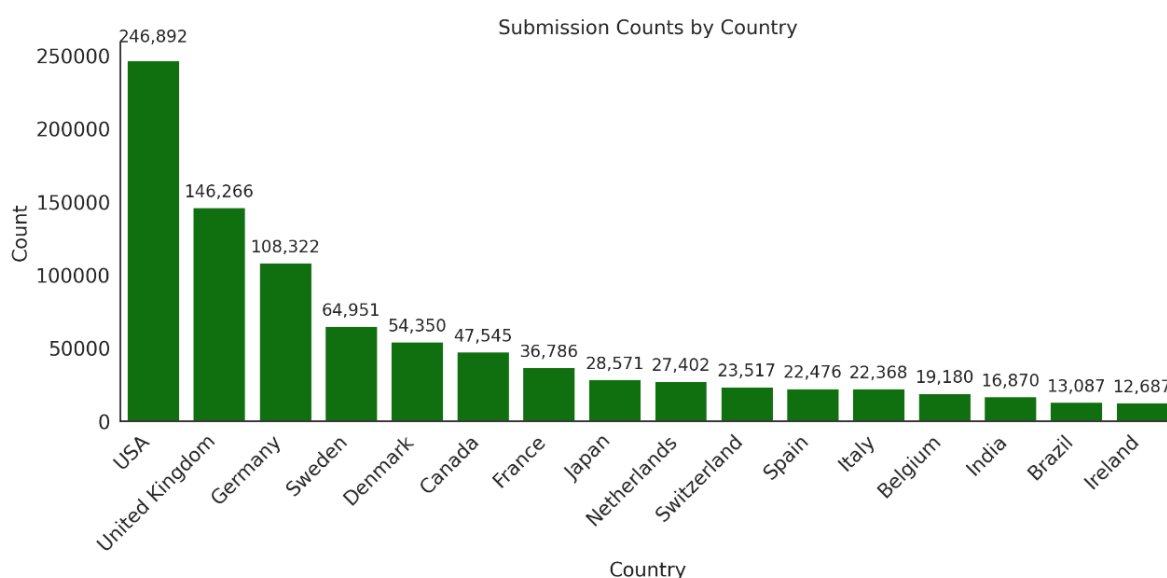


Figure 2.16. Sequence submissions by different countries on GISAID database where y-axis shows the number of sequences, and x-axis shows the country.

2.9.5. Variant distribution in NCBI dataset

The variant distribution across the NCBI sequence dataset showed that the majority of submissions belonged to the delta variant, with a total of 606,385 sequences. The second-highest submissions belonged to non-VOCs, with 176,057 sequences. The Beta variant had the fewest submissions, totalling 2,379, as shown in Table 2.1. and Figure 2.17.

Variants	Total number of sequences in the Feb NCBI ~1 million seq dataset	Colour code
Alpha B.1.1.7 and subvariants	120,808	
Beta B.1.351 and subvariants	2,379	
Gamma P.1, P.2 and subvariants	11,702	
Delta B.1.617.2, AY., XC and subvariants	606,385	
Omicron BA.1 and subvariants	153,420	
Non-VOC	176,057	
Total	1,070,751	

Table 2.1. Distribution of sequences belonging to Alpha (pink), Beta (dark blue), Gamma (cyan), Delta (yellow), Omicron (orange) and Non-VOCs (purple) SC-2 variants in the NCBI dataset.

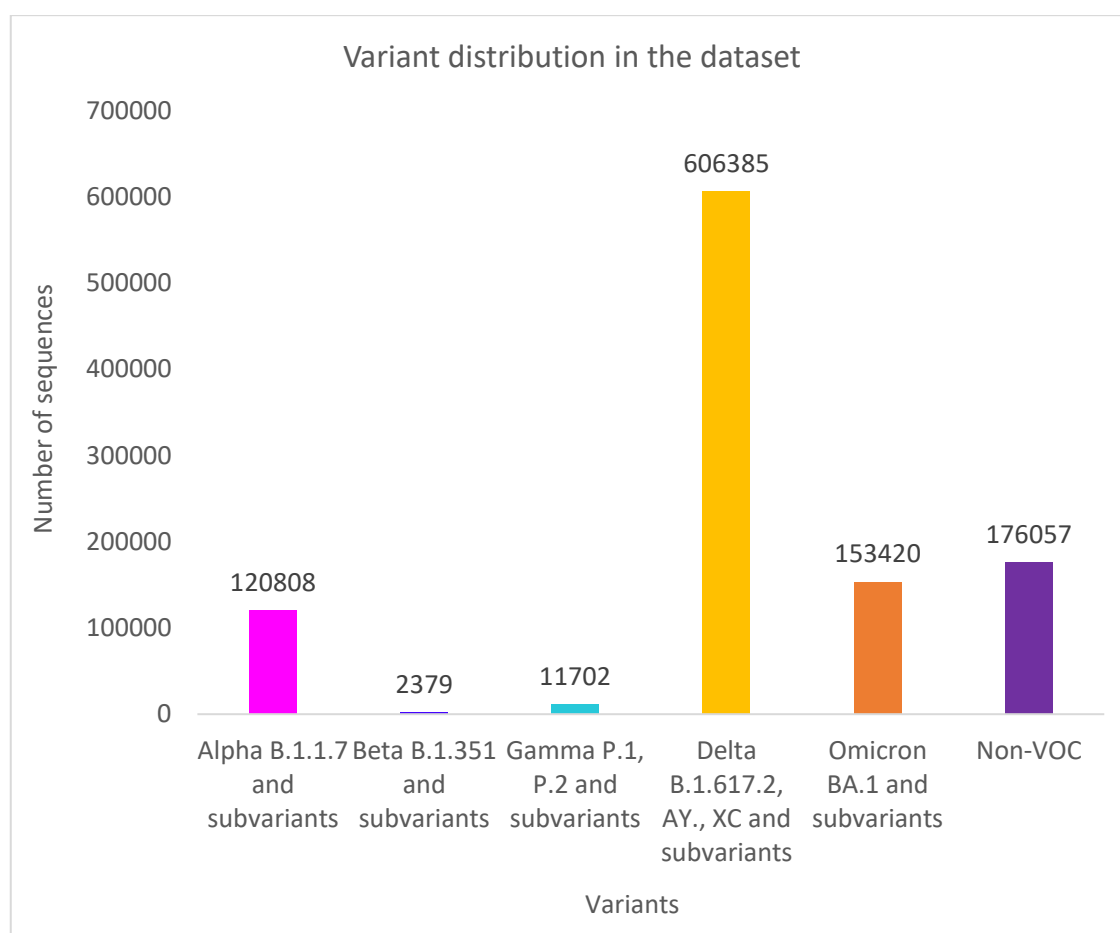


Figure 2.17. Bar graph of sequence distribution in the NCBI dataset. Alpha (pink) , Beta (dark blue), Gamma (cyan), Delta (yellow), Omicron (orange) and Non-VOCs (purple). The x-axis shows the number of sequences and y-axis shows the SC-2 variants.

2.9.6. NSP conservation/mutational analysis

The MSA output of the SC-2 ORF1ab protein sequences from the NCBI dataset consisting of **1,070,751** sequences was chopped into NSP1-NSP16 alignments. For each NSP, conservational and mutational analysis was performed, and the conserved/mutated amino acid residues were then visualised on a 3D-protein structure. For the visualisation, PDB submissions, loop modelled structures with MOE and AlphaFold (Jumper et al., 2021) predictions were used (Table 2.2.).

Protein	Actual length	PDB available	Protein length available	Structure used
NSP1	1–180 (180 length)	Total 25 structures (N and C terminal separate)	112 (resolution 1.65 Å)	7K3N (N-terminal) + loop modelling using MOE
NSP2	181–818 (638 length)	Total 3 structures	638 (resolution 3.76 Å)	7MSW only one with complete structure
NSP3	819–2763 (1945 length)	Total 114 structures Mac1 domain	169 (resolution 0.88 Å)	7KQP (adp co-crystalised)
NSP4	2764–3263 (500 length)	Total 3 structures	463 (resolution 4.60 Å)	8YB7 Electron microscopy + loop modelling using MOE
NSP5	3264–3569 (306 length)	Total 602 structures	306 (resolution 1.74 Å)	8UH5 X-ray
NSP6	3570–3859 (290 length)	0	0	AlphaFold
NSP7	3860–3942 (83 length)	Total 24 structures	75 (resolution 2.57 Å)	7DCD + loop modelling using MOE
NSP8	3943–4140 (198 length)	Total 31 structures	115 (resolution 2.50 Å)	6M5I but it had too many proteins to build, slow loop modelling in MOE so instead used AlphaFold
NSP9	4141–4253 (113 length)	Total 12 structures	104 (resolution 2.12 Å)	8WKE + loop modelling using MOE

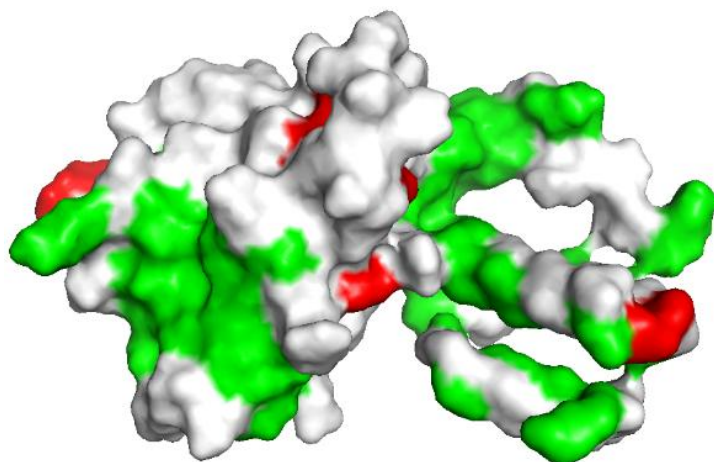
NSP10	4254–4392 (139 length)	Total 8 structures	126 (resolution 2.39 Å)	8VUO + loop modelling using MOE
NSP11	4393–4405 (13 length)	NA	NA	NA
NSP12	4393–5324 (932 length)	Total 51 structures	932 (resolution 2.58 Å)	9DJ8 (electron microscopy) as it was complete
NSP13	5325–5925 (601 length)	Total 118 structures	601 (resolution 2.43 Å)	5RML
NSP14	5926–6452 (527 length)	Total 63 structures	501 (resolution 1.82 Å)	5SLP initial amino acids were missing and not modelled so instead used AlphaFold
NSP15	6453–6798 (346 length)	Total 47 structures	346 (resolution 3.43 Å)	7TQV (electron microscopy)
NSP16	6799–7096 (298 length)	Total 28 structures	298 (resolution 2.05 Å)	7JYY

Table 2.2. NSP sequence lengths along with the type of structure submission (PDB, AlphaFold, loop modelled) used for 3D visualisation. The table shows the protein's name, its actual length (total number of amino acid residues in the native protein), the number of resolved structures available in the PDB, the portion of the protein for which structural data are available, and the specific PDB ID or modelled structure used in this study.

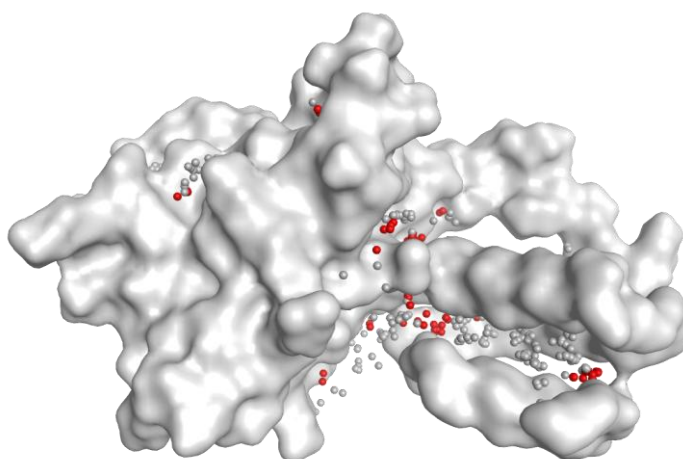
2.9.7. NSP1

For SC-2 NSP1, 7K3N (N-terminal) PDB was used and rest of the amino acids of the C-terminal was modelled using MOE loop modelling. The top 10 mutations/deletions in the SC-2 NSP1 sequence dataset were MET85-, PHE143-, SER142-, LYS141-, VAL84-, ARG24CYS, HIS110TYR, HIS83-, GLU87ASP and GLY82-. The mutational and conservational threshold selected were 0.237% and 0.012% respectively. The amino acid positions above the mutational threshold are highly variable whereas the amino acid positions below the conservational threshold are highly conserved. The amino acid positions present between these thresholds have intermediate conservation (Figure 2.18. a). As per site finder in MOE, a total of 8 binding pockets (Figure 2.18. b,c,d) were identified and among them site 1 was the largest binding pocket consisting of conserved amino acid residues. The second most conserved binding pocket was site 6.

(a)



(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	261	2.61	28	53	1: (GLU2 SER3 LEU4 PHE8 ASN9 GLU10 LYS11 HIP13 VAL14 GLU36 LEU39 SER40 ARG43 GLN44 LYS47 VAL86 GLU87 LEU88 LEU123 ARG124 LYS125 ASN126 GLY127 ASN128 LYS129 GLY130 ALA131 GLY132 GLY133 HIS134 TYR136 GLY137 ALA138 LYS141 LEU145 GLY146 GLU148 LEU149 GLY150 PRO153 TYR154 ASP156 PHE157 ASN160 TRP161 ASN162 THR163 LYS164 HIS165 SER166 GLY168 VAL169 THR170 GLU172 LEU173 MET174 GLU176 LEU177 ASN178 GLY179)
2	17	-0.00	8	17	1: (ARG73 ASP75 ARG77 THR78 ALA79 HIS81 GLY82 HIS83 VAL84 MET85)
3	26	-0.24	8	22	1: (VAL14 SER74 MET85 VAL86 GLU87 ARG99 SER100 LEU122 LEU123 ARG124)
4	17	-0.44	8	16	1: (GLU55 VAL56 GLU102 THR103 LEU104 GLY105 ARG119)
5	18	-0.46	10	12	1: (VAL20 LEU21 GLN22 VAL23 VAL54 GLU55 VAL60 VAL116)
6	14	-0.47	4	6	1: (HIP13 VAL14 LEU123 ARG124 LYS125 ASN126 GLY127 ASN128)
7	5	-0.50	4	7	1: (THR12 GLN15 LEU16 SER17 LEU18 GLY49)
8	10	-0.50	4	7	1: (VAL28 GLY30 PHE31 GLY32 ASP33 GLU65 GLN66 PRO109)

(d)

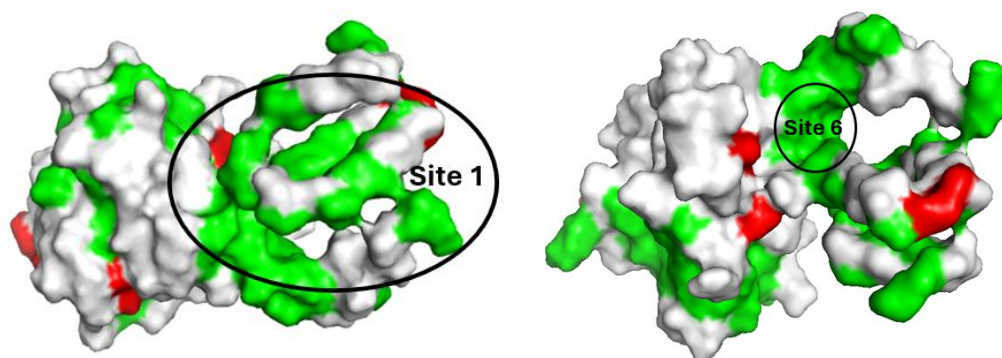


Figure 2.18. Conservational and mutational analysis of NSP1 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (e) Visualisation of second most conserved binding pockets, site 1 and 6.

There are no co-crystal structures with a known inhibitor, but an *in-silico* study conducted by Afsar et al. (2022) found that montelukast sodium hydrate might bind to NSP1, thereby inhibiting its interaction with the 40S ribosomal subunit. It was found that montelukast had a binding affinity (K_d) of $10.8 \pm 0.2 \mu\text{M}$ *in-vitro* and also showed antiviral activity against SC-2 with reduced viral replication in HEK cells expressing ACE2 and Vero-E6 cells. The binding site in which montelukast was docked consisted of amino acid residues that are a part of site 1 in our study. These amino acids were GLU148, LEU149, PRO153, PHE157, ASN160, LYS164, HIS165, ARG171, and LEU173 (Figure 2.19.). Among them, LEY149, PHE157, LYS164 and LEU173 were highly conserved in our analysis. Whereas GLU148, PRO153, ASN160, HIS165 and ARG171 had intermediate conservation. As montelukast might bind to site 1 of NSP1, which is a conserved binding pocket across the SC-2 variants in our analysis, it has the potential to be used as a pan SC-2 inhibitor.

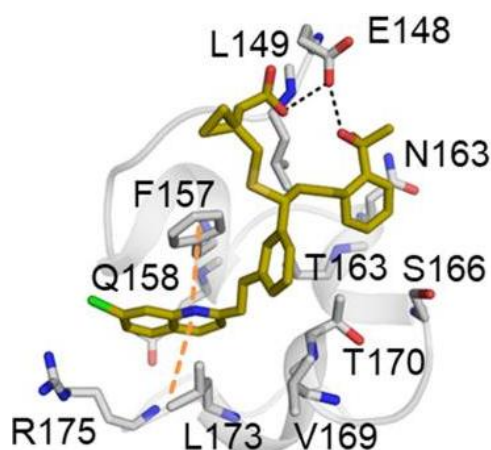
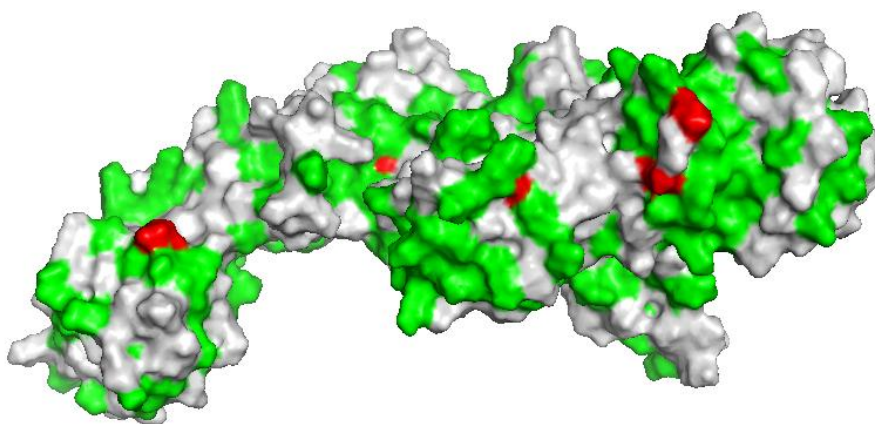


Figure 2.19. Binding mode of montelukast (yellow) with NSP1 (grey) after 500ns MD run. Hydrogen bonds are represented in dotted lines. Schematic taken from Afsar et al., 2022.

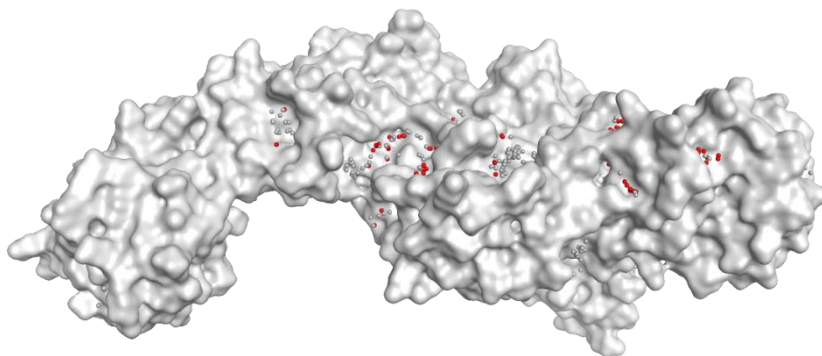
2.9.8. NSP2

The only complete structure of NSP2 was the PDB ID: 7MSW. There is currently no known inhibitor of the NSP2 protein. The top 10 mutations in the SC-2 NSP2 sequence dataset were THR85ILE, PRO129LEU, GLY339SER, GLU587LYS, LYS81ASN, ALA419VAL, ILE120PHE, TRP466LEU, SER358ALA and THR223ILE. The mutational and conservational threshold selected were 0.547% and 0.013% respectively (Figure 2.20. a). As per site finder in MOE, a total of 31 binding pockets (Figure 2.20. b,c,d) were identified and among them sites 1, 2, 3, 6 and 16 had a greater number of conserved amino acid residues (Figure 2.20. c,d). These conserved pockets can be targeted for the identification of inhibitors.

(a)



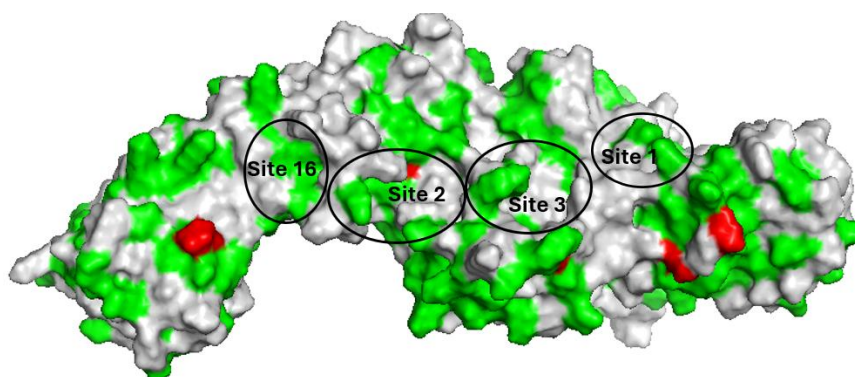
(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	43	2.53	11	25	1: (GLU89 CYS90 PRO91 ASN92 PHE93 PHE95 PRO96 LEU97 VAL258 VAL259 GLY260 GLU261 GLY262 SER263 GLU264 ASN267)
2	64	2.47	18	49	1: (VAL284 GLY285 ASP286 LYS288 ASN290 GLU291 GLY344 GLN346 ASP409 LEU410 ALA411 THR412 ASN413 ARG463 TRP466 GLU467 LYS470 PHE471 THR474 LYS489 GLU490 LYS492)
3	66	1.50	22	40	1: (LEU266 ASP268 ASN269 LEU270 GLU272 ILE273 GLU292 ALA340 TRP341 ASN342 LYS347 SER348 SER351 PRO352 LEU353 TYR354)
4	40	1.31	17	33	1: (PHE163 GLU291 GLN346 LYS347 SER348 ILE349 LEU350 PHE368 SER369 ARG370 THR371 THR412 ASN413 ASN414)
5	18	1.09	10	20	1: (CYS136 SER138 TRP151 GLN152 THR153 THR176 ILE349 LEU350 SER351 PRO352 LEU353 TYR354 ALA355 ARG365)
6	69	1.02	22	34	1: (ASN133 TYR179 LEU180 PRO181 GLN182 CYS253 ASN254 SER299 PHE300 SER301 ALA302 ASN328 PHE329 PHE356 SER358 GLU359)
7	26	0.87	12	19	1: (ASN92 PHE93 VAL94 LYS171 GLU172 GLY173 ALA174 THR175 ALA225 PHE226 GLY227 GLY228 CYS229 VAL259 GLY260)
8	31	0.72	16	21	1: (GLN383 ASP402 PHE406 TYR420 ILE421 GLY424 VAL425 GLN427 LEU428 VAL469)
9	10	0.45	5	9	1: (GLY516 GLY517 ALA518 GLY581 ASP582 LEU583 ASN606)
10	33	0.32	12	14	1: (ASN435 ILE436 CYS477 GLU478 ILE479 LYS534 GLY535 LEU536 GLY564 GLU565)
11	24	0.19	5	14	1: (ASN130 CYS132 ASN133 GLN134 LYS187 ALA302 ALA357 SER358 GLU359 ARG362)
12	27	0.16	13	31	1: (SER65 LYS67 PHE75 GLU76 ILE77 LYS78 THR103 ILE104 GLN105 PRO106 ARG246)
13	28	0.02	11	21	1: (ASP14 TYR16 PRO17 GLU19 LYS22 GLU70 LEU71 GLN72 THR73 ARG218 LYS219 GLY220)
14	11	-0.03	12	16	1: (VAL48 GLU57 ILE58 PHE83 THR85 PHE86)
15	23	-0.19	11	19	1: (LYS102 THR103 ILE104 TYR124 LEU180 PRO181 GLN182 ASN183 ALA184 PRO245 ARG246 ALA247 SER248)
16	17	-0.32	15	20	1: (ILE484 VAL485 THR486 ILE491 GLN496 PHE499 LYS500 ASN503)
17	22	-0.40	9	19	1: (LEU204 HIS208 GLY220 GLY221 ARG222 THR223 SER232 TYR233 ARG246)
18	11	-0.41	5	10	1: (THR569 GLU570 VAL571 ASN625 MET626 MET627 VAL628)
19	28	-0.60	5	13	1: (PHE116 CYS193 PRO200 GLU201 HIS202 CYS236 HIS237 ASN238 LYS239)
20	19	-0.62	13	20	1: (ILE104 PRO106 ARG119 SER122 VAL123 SER232 TYR233 VAL234 VAL244 ARG246)
21	25	-0.71	9	22	1: (THR153 GLY154 THR160 GLU162 ARG362 VAL363 SER366 ARG370 ILE393)
22	20	-0.75	7	12	1: (ARG4 TYR5 PRO13 LEU18 GLU19 CYS20)
23	13	-0.77	3	15	1: (HIS208 ASN209 GLY212 LEU213 THR215 ARG222)
24	7	-0.80	3	5	1: (PHE287 GLY313 LEU314 ASP315 THR407 SER408 ASP409)
25	17	-0.82	5	9	1: (CYS164 GLY165 ARG370 ILE389 GLY392 ILE393)
26	9	-0.91	9	14	1: (LEU24 LYS45 ARG46 TYR49 CYS50 ARG52)
27	10	-0.97	8	12	1: (LYS81 LYS82 LEU97 ASN98)
28	21	-1.09	3	4	1: (PHE156 THR170 LYS171 ILE224 ALA225 PHE226)
29	15	-1.09	4	4	1: (PRO13 ASP14 ASN92 PHE93 VAL94)
30	6	-1.09	1	3	1: (GLU612 ILE613 LYS614 ASP615 THR616)
31	15	-1.09	4	9	1: (ILE100 ASN250 ILE251 GLY252 GLN275 LYS276 GLU277)

(d)



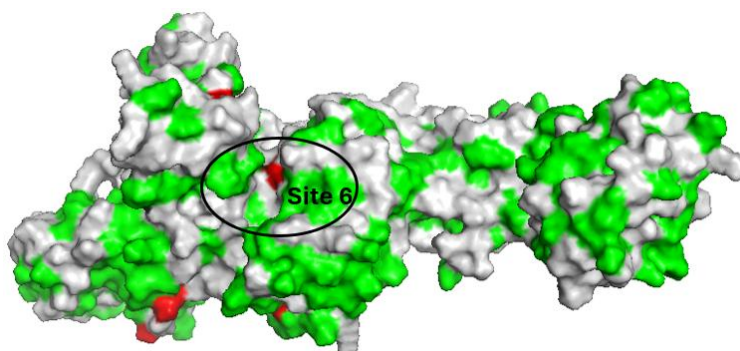


Figure 2.20. Conservational and mutational analysis of NSP2 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding pockets, sites 1,2,3,6 and 16.

2.9.9. NSP3 (Mac1 domain)

With regards to this project, we have worked on the Mac1 domain of the NSP3 protein. There are many PDB's available with lead-like compounds co-crystallized within the binding pocket of NSP3 Mac1 domain which are discussed in Chapter III. For visualisation of the conserved/mutated residue 7KQP with ADP in the binding pocket was used. The top 10 mutations in the SC-2 NSP3 Mac1 sequence dataset were SER164LEU, ASP103TYR, THR31ALA, LYS74ASN, PRO134SER, SER5GLY, LEU106PHE, VAL32LEU, THR144ILE and ALA25VAL. The mutational and conservational threshold selected were 0.171% and 0.013% respectively (Figure 2.21. a). Using site finder in MOE, a total of 10 binding pockets (Figure 2.21. b,c,d) were identified and among them sites 1 and 2 were the largest binding sites with a greatest number of conserved amino acid residues (Figure 2.21. c,d).

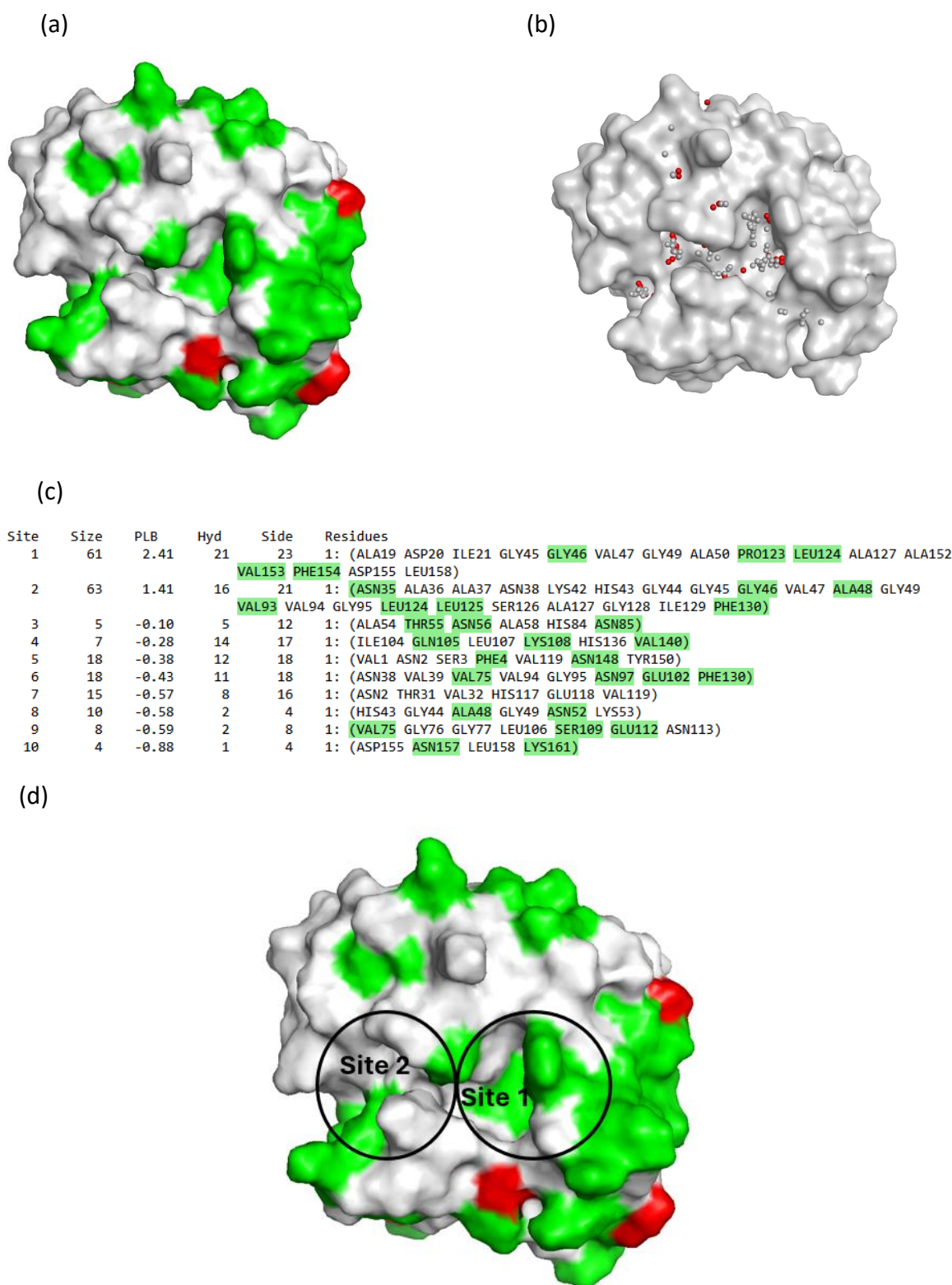


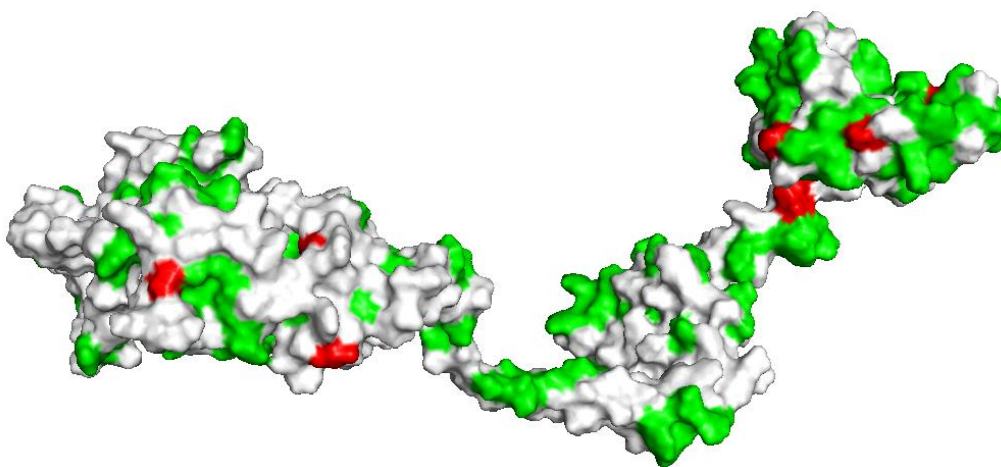
Figure 2.21 Conservational and mutational analysis of NSP3 Mac1 domain (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also

resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hvd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding pocket, sites 1 and 2.

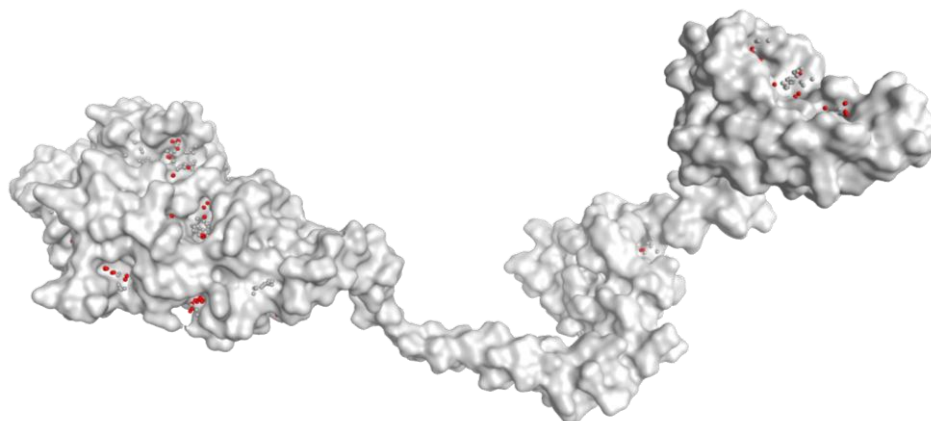
2.9.10. NSP4

The PDB ID 8YB7 with a loop modelled structure was used for the visualisation of the conserved/mutated residues. There is currently no known inhibitor of the NSP2 protein. The top 10 mutations in the SC-2 NSP4 sequence dataset were VAL167LEU, THR492ILE, ALA446VAL, LEU438PRO, ALA92VAL, ASP217ASN, LYS399GLU, PHE375SER, SER395THR and THR204ILE. The mutational and conservational threshold selected were 0.540% and 0.004% respectively (Figure 2.22. a). Using site finder in MOE, a total of 32 binding pockets (Figure 2.22. b,c,d) were identified and among them site 7 (THR411, PHE412, GLU413, ALA416, LEU435, GLN440, TYR444, ARG464, GLU465, ALA467, CYS468) was the binding site which was entirely conserved (Figure 2.22. c,d). This site could be interesting to target and identify inhibitors.

(a)



(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	102	3.16	33	64	1: (THR135 PRO136 SER137 LYS138 LEU139 ILE140 GLU141 TYR142 THR143 THR147 SER148 ALA149 CYS150 VAL151 TYR182 GLU183 SER184 LEU185 ARG186 PRO187 PHE201 PRO202 ASN203 THR204 GLU207 GLY208 SER209 ARG211 VAL213 SER238 GLY239 ARG240 TRP241)
2	105	2.95	39	65	1: (TRP6 ILE11 LEU18 PHE19 VAL20 ALA21 ALA22 ILE23 VAL30 HIS31 VAL32 MET33 SER34 LYS35 PHE64 ALA65 ASN66 LYS67 ASN84 ASP85 LYS86 LEU90 ARG112 ASN115 GLY116 ASP117 PHE118 HIS120)
3	111	2.48	41	55	1: (THR54 GLU98 VAL102 VAL103 LEU106 PRO107 GLY108 THR109 LEU122 PRO123 ARG124 VAL125 PHE126 ALA128 VAL129 GLY130 ASN131 ILE132 TYR134 LEU139 ILE140 GLU141 TYR142 ALA153 ALA154 GLU155 TYR171)
4	50	0.82	17	29	1: (TYR448 TYR451 LYS452 PHE454 SER455 GLY456 SER462 TYR463 ALA466 ALA467 HIS470 PRO490 GLN491 THR492 SER493 LEU499)
5	60	0.80	17	34	1: (GLU230 ALA231 ASN244 ARG249 SER250 LEU251 PRO252 GLY253 VAL254 PHE255 CYS256 GLY257 VAL258 LEU263 ASN266 MET267 THR269 PRO270)
6	17	0.36	12	14	1: (PHE159 LYS160 ASP161 VAL167 TYR169 ARG190 TYR191 VAL192)
7	29	0.25	18	30	1: (THR411 PHE412 GLU413 ALA416 LEU435 GLN440 TYR444 ARG464 GLU465 ALA467 CYS468)
8	21	-0.08	12	31	1: (ASN4 ASN5 LEU7 LYS35 HID36 THR37 ASP38 ALA69 THR113 THR114 ASN115)
9	40	-0.10	25	33	1: (LEU18 PHE19 VAL20 ALA21 VAL103 PRO104 GLY105 ILE158 PRO168 LEU176 GLU177 LEU193 MET194 ASP195)
10	11	-0.15	4	12	1: (SER197 ILE198 ILE199 CYS221 ARG222 HIS223 GLY224 THR225 CYS226)
11	23	-0.19	14	28	1: (PHE145 ALA146 THR147 ARG222 HIS223 THR225 ARG240 TRP241 TYR247 TYR248 LEU251)
12	4	-0.21	8	17	1: (SER236 THR237 SER238 ARG240 LEU251 PRO252 GLY253 VAL254)
13	19	-0.23	15	21	1: (ILE49 LEU90 LEU106 PRO107 GLY108 THR109 ILE110 HIS120 PHE121)
14	26	-0.23	10	18	1: (GLY52 GLY105 LEU106 PRO107 LEU152 ALA153 ALA154 CYS156 THR157 LEU193 ASP195 HIS223)
15	29	-0.24	14	26	1: (THR350 PHE351 LEU353 ASN355 SER358 PHE359 HIS362 ILE363 CYS384 ILE385 LYS388)
16	28	-0.30	9	16	1: (HIS470 LEU471 LYS473 ALA474 ASP477 LEU486 TYR487 GLN488 PRO489 PRO490 GLN491)
17	12	-0.30	14	17	1: (LEU15 VAL16 PHE17 LEU18 PHE101 VAL103 ILE110 ARG112 PHE118)
18	11	-0.49	10	18	1: (ASP50 GLY51 VAL53 THR54 ARG55 PHE145 HIS223 GLY224)
19	5	-0.52	3	7	1: (VAL407 SER408 PHE409 MET426 LYS429 LEU430)
20	9	-0.53	7	10	1: (SER137 LYS138 LEU139 TYR171)
21	7	-0.53	2	3	1: (SER336 PHE337 LEU338 PRO339 GLY340 VAL341)
22	11	-0.54	8	12	1: (ALA474 ASP477 PHE478 SER481 SER483 VAL485)
23	19	-0.56	6	11	1: (LYS1 ILE2 VAL3 ASN4 ASN5 VAL32 MET33 SER34 LYS35)
24	9	-0.57	6	6	1: (PHE101 VAL102 VAL175)
25	14	-0.57	3	5	1: (ILE49 ASP50 GLY51 GLY52 ALA87 CYS88 PRO89 LEU90)
26	6	-0.57	8	12	1: (SER40 ILE43 ILE44 GLY45 TYR46 ASP70 THR73 TRP74)
27	15	-0.60	11	22	1: (ASP188 ARG190 GLN200 PHE216 SER218 GLU219 TYR220)
28	16	-0.63	25	25	1: (PRO136 VAL151 LEU152 PHE159 TYR182 LEU185 TYR191 ILE199 PHE201)
29	10	-0.63	8	13	1: (GLU98 VAL99 GLY100 PHE101 VAL102 THR109 PRO123 GLU155 VAL175)
30	14	-0.64	4	4	1: (THR215 ALA231 GLY232 VAL233 VAL258 ASP259 ALA260)
31	11	-0.71	3	10	1: (LYS452 HIS470 PRO490 GLN491 THR492 SER493)
32	9	-0.71	5	7	1: (PHE304 VAL315 ALA316 THR319 LEU320)

(d)

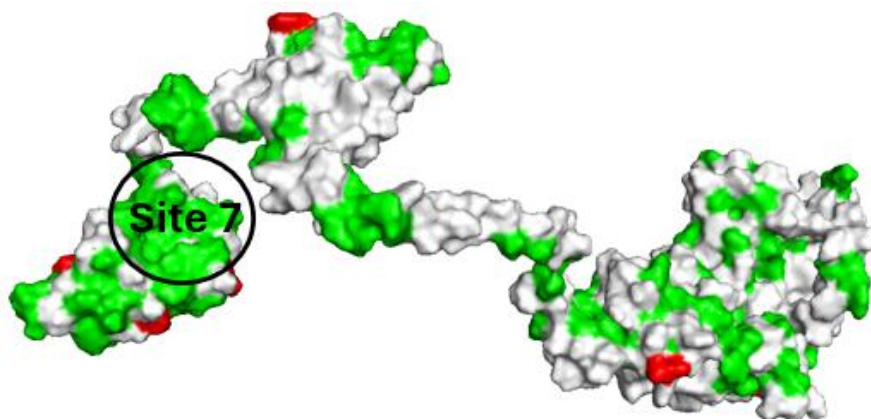
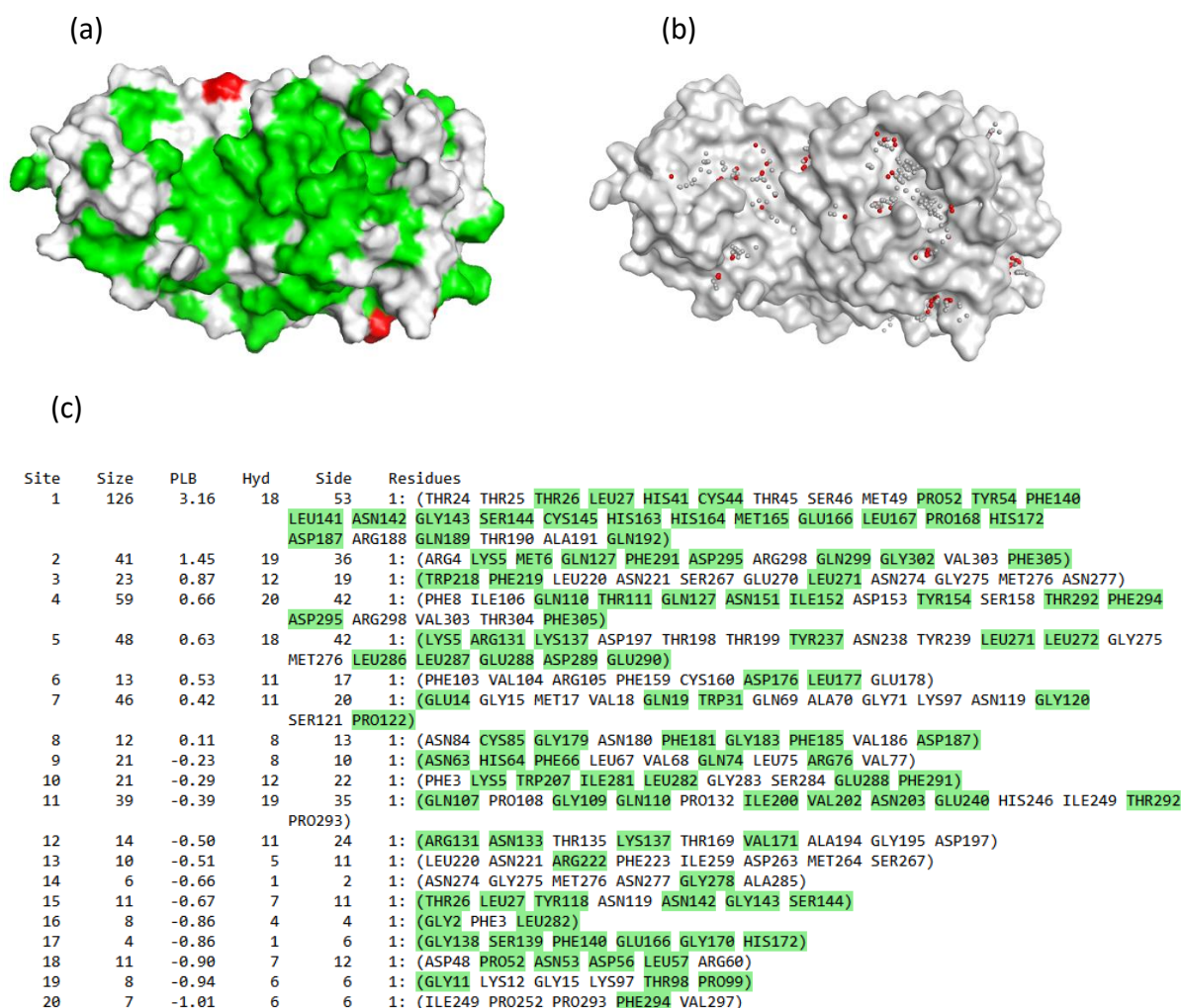


Figure 2.22. Conservational and mutational analysis of NSP4 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity

(Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding pocket, site 7.

2.9.11. NSP5

The PDB ID 8UH5 co-crystallised with an inhibitor TKB-272 (Higashi-Kuwata et al., 2025) was used for the visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP5 sequence dataset were PRO132HIS, LEU89PHE, LYS90ARG, VAL73ILE, ILE259LEU, ALA260VAL, THR196MET, PRO96LEU, PRO241SER and PRO108SER. The mutational and conservational threshold selected were 0.240% and 0.003% respectively (Figure 2.23. a). As per site finder in MOE, a total of 20 binding pockets (Figure 2.23. b,c,d) were identified and among them sites 1, 4 and 5 had greater number of conserved amino acid residues (Figure 2.23. c,d). The inhibitor TKB-272 binds to site 1 of NSP5 (Figure 2.23. e) which is a conserved binding pocket across the SC-2 variants in our analysis, it has the potential to be used as a pan SC-2 inhibitor. Interestingly, FDA authorized Paxlovid (nirmatrelvir) also binds to site 1 of NSP5.



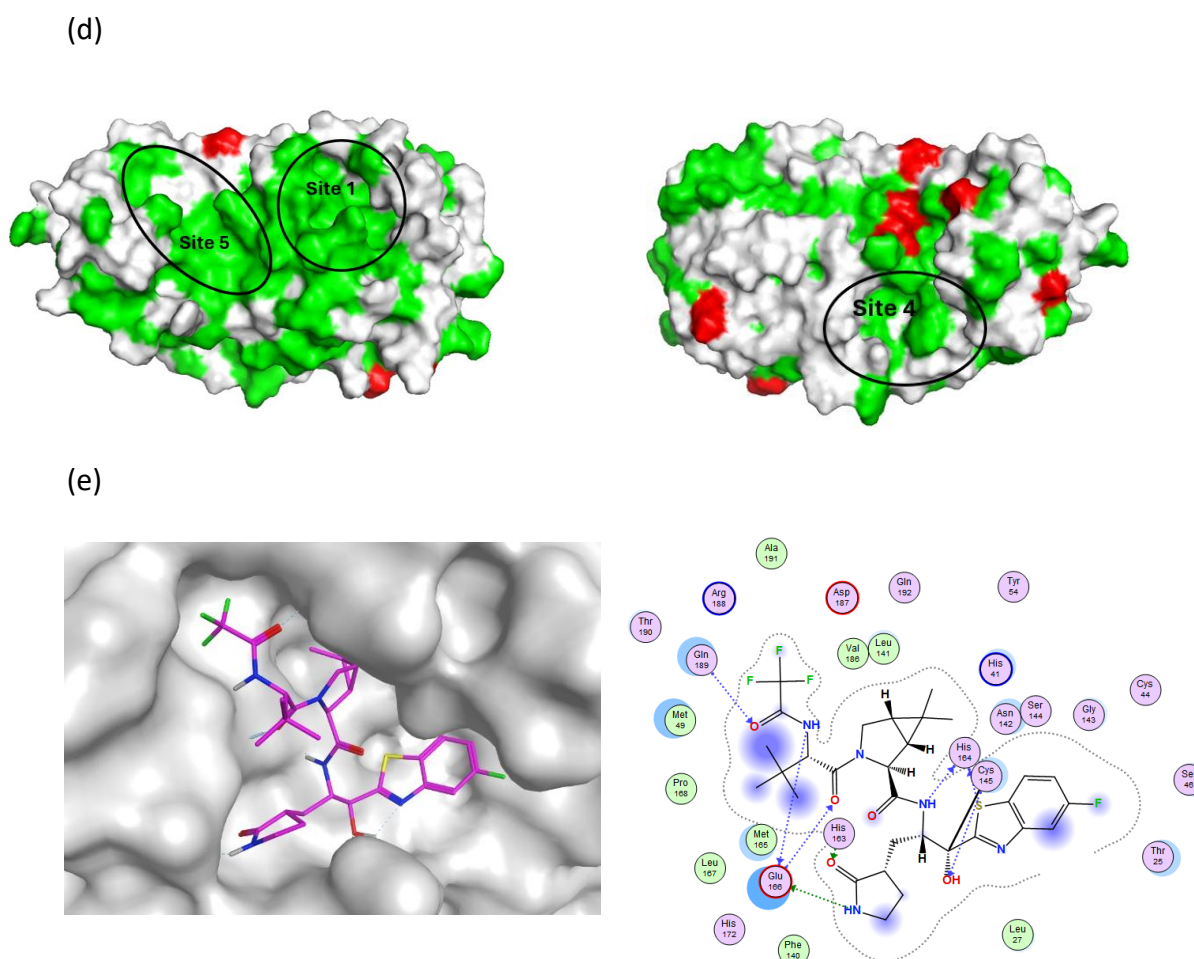


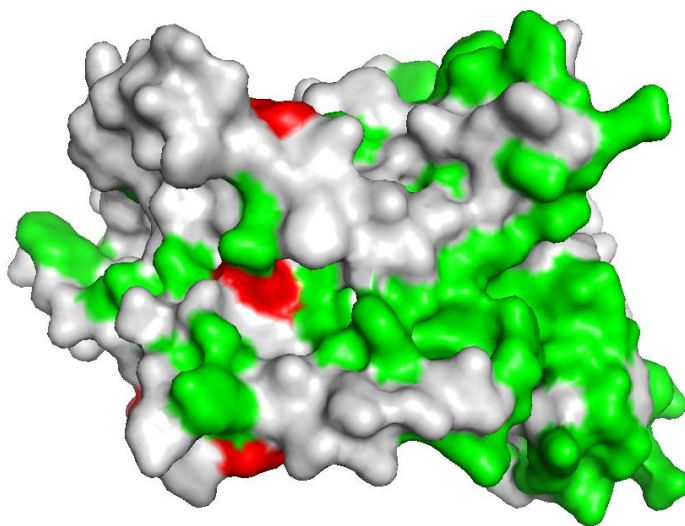
Figure 2.23. Conservational and mutational analysis of NSP5 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding sites 1,4 and 5. (e) TKB-272 (pink) inhibitor within the site 1 binding pocket (grey) with its protein interaction map. Hydrogen bonds are depicted in dotted lines.

2.9.12. NSP6

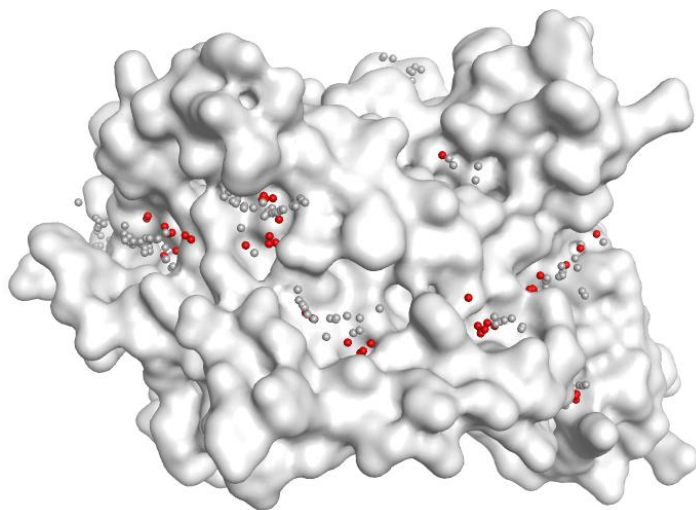
There were no complete structures of NSP6, hence a predicted structure from AlphaFold was used for the visualisation of the conserved/mutated residues. The top 10 mutations/deletions in the SC-2 NSP6 sequence dataset were THR103-, SER104-, SER106-, GLY107-, PHE108-, THR77ALA, LEU105PHE, ILE189VAL, VAL149ALA and ILE162VAL. The

mutational and conservational threshold selected were 7.223% and 0.005% respectively (Figure 2.24. a). As per site finder in MOE, a total of 20 binding pockets (Figure 2.24. b,c,d) were identified and among them sites 13 and 16 had highest number of conserved amino acid residues (Figure 2.24. c,d). Currently, there are no known inhibitor for NSP6, thus these conserved binding pockets can be explored for the identification of novel inhibitors.

(a)



(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	122	2.89	27	61	1: (SER1 ALA2 LYS4 ARG5 THR6 LYS8 GLY9 THR10 HIS12 TRP13 MET58 LYS61 HIS62 LYS63 HIS64 VAL241 LEU260 PRO261 PRO262 LYS263 ASN264 SER265 ASP267 THR288 VAL289 GLN290)
2	31	1.49	13	24	1: (ILE124 THR127 ALA128 VAL131 TYR132 ASP133 ARG137 VAL171 THR172 LYS281 PRO282)
3	77	1.06	26	49	1: (GLN30 TRP31 SER32 PHE35 PHE81 ASN82 TYR85 MET86 PRO87 TRP90 VAL91 MET95 LYS109 ASP112 THR203 GLY204 ASN205 THR206 GLN208)
4	89	1.00	34	60	1: (LEU22 LEU25 VAL26 THR29 GLN30 TYR38 PHE42 ASN82 ILE94 MET95 LEU98 MET100 VAL101 ASP102 THR103 LYS109 LEU110 LYS111 ASP112 MET115)
5	63	0.90	34	42	1: (THR20 LEU23 VAL24 PHE45 GLY48 ILE49 MET52 SER53 ALA56 MET57 PHE66 LEU71 LEU212 VAL213 CYS215 PHE216 TYR219)
6	33	0.80	19	31	1: (THR130 TYR132 ASN174 TYR175 THR238 VAL241 TYR242 ASP243 LEU245 LYS285 THR288)
7	21	-0.15	23	33	1: (LYS8 HIS11 LEU14 LEU15 ILE18 LEU19 LEU122 ILE266 ASP267 LYS270)
8	37	-0.21	18	25	1: (LEU158 ILE162 ARG187 VAL190 PHE191 VAL194 THR206 LEU207 ILE210)
9	33	-0.26	23	28	1: (TYR38 GLU39 PHE42 ARG93 ILE94 TRP97 LEU98)
10	23	-0.29	9	23	1: (ASN144 THR147 LEU148 LYS151 ASP159 ILE162 SER163 ALA166 LEU167 ARG187)
11	6	-0.40	8	17	1: (ASP133 ARG137 ARG138 THR141 VAL171)
12	22	-0.45	15	20	1: (LEU19 LEU22 VAL26 LEU75 LYS111 VAL114 MET115 SER118 ALA119 LEU122)
13	15	-0.60	14	26	1: (TYR132 ASP134 ALA136 ARG137 TRP140 SER170 SER173 ASN174 SER176 GLY177)
14	26	-0.61	14	19	1: (LYS8 GLY258 LEU259 LEU260 PRO261 PRO262 ASP267 LYS270 LEU271 LYS274)
15	6	-0.62	2	10	1: (LYS61 HIS62 PHE228 CYS229 ASN232 ARG233)
16	5	-0.85	3	3	1: (ALA51 ALA54 PHE55)
17	5	-0.93	7	9	1: (VAL178 THR181 VAL182 CYS221 LEU227)
18	8	-0.93	2	5	1: (PHE59 TYR219 CYS223)
19	8	-0.93	6	10	1: (CYS193 CYS197 PHE200 PHE201)
20	3	-0.93	4	6	1: (TYR80 MET83 VAL84 PHE108 CYS113)

(d)

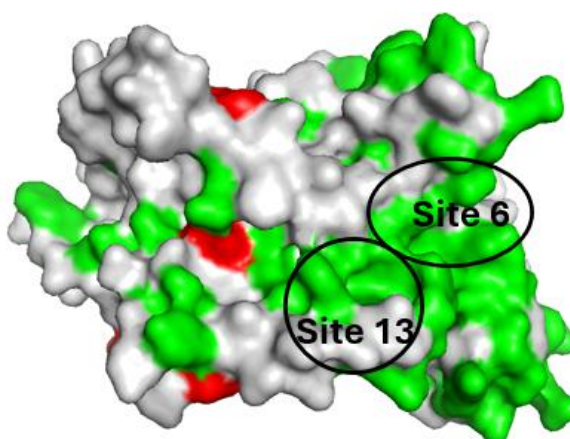


Figure 2.24. Conservational and mutational analysis of NSP6 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding sites 13 and 16.

2.9.13. NSP7

The PDB ID 7DCD with a loop modelled structure was used for the visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP7 sequence dataset

were LEU71PHE, ASP77ASN, SER25LEU, MET75ILE, GLN63ARG, SER26PHE, THR81ILE, ASP44ALA, VAL33ILE and MET3ILE. The mutational and conservational threshold selected were 0.058% and 0.003% respectively (Figure 2.25. a). As per site finder in MOE, a total of 4 binding pockets (Figure 2.25. b,c,d) were identified and among them site 2 had greatest number of conserved amino acid residues (Figure 2.25. c,d). Currently, there are no known inhibitor for NSP7, thus the conserved binding pocket can be explored for the identification of novel inhibitors.

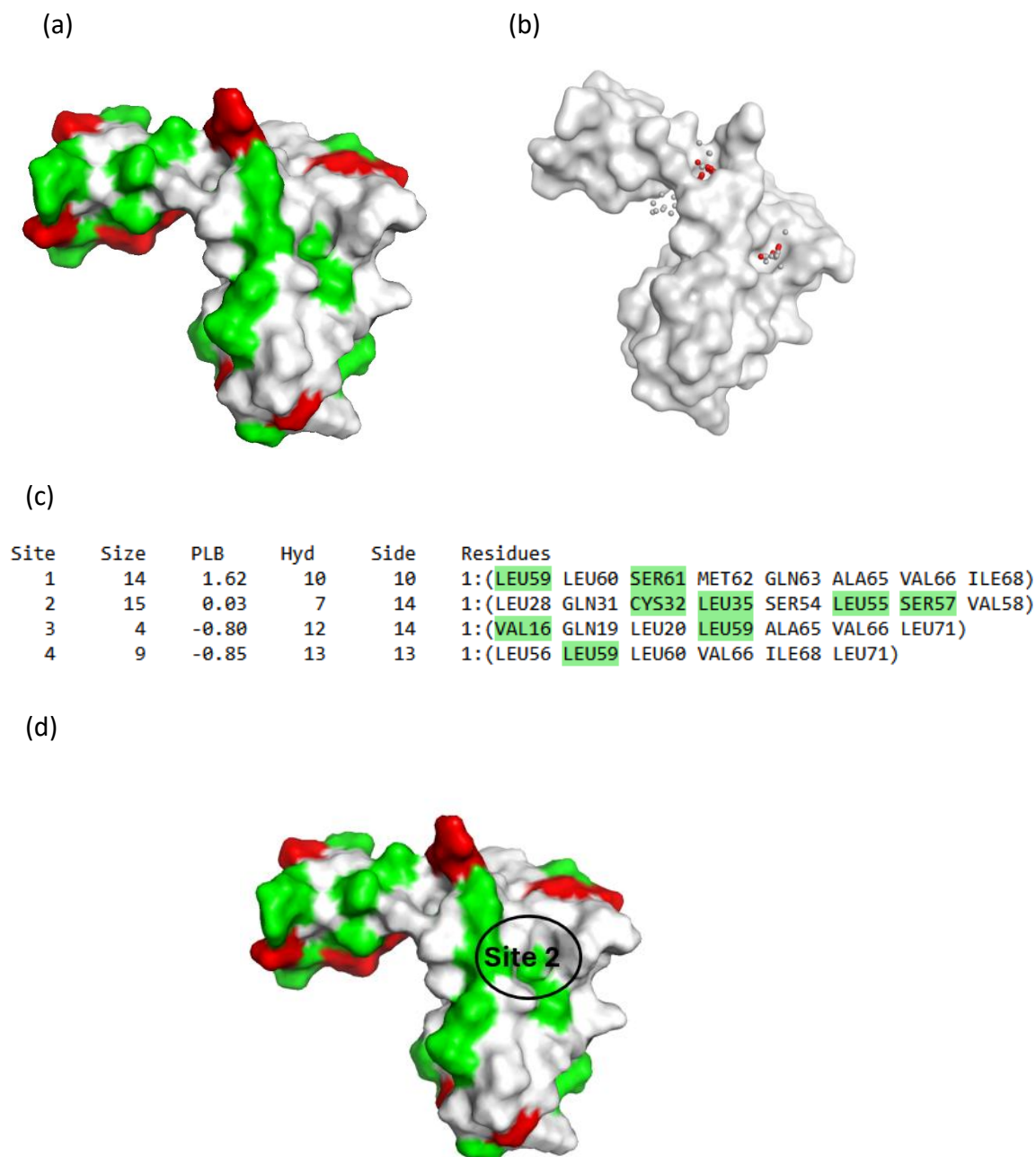


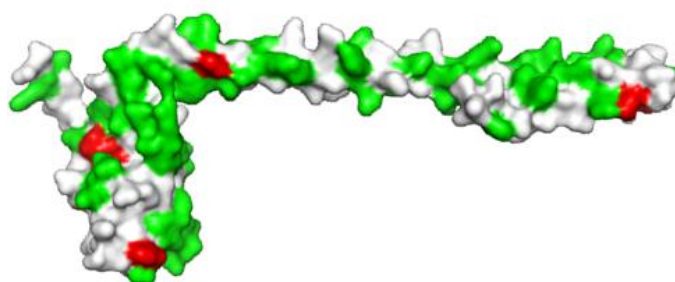
Figure 2.25. Conservational and mutational analysis of NSP7 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid

residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding site 2.

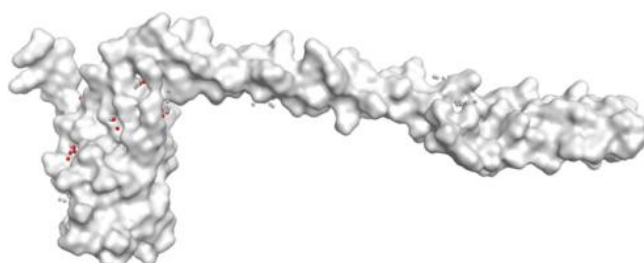
2.9.14. NSP8

The AlphaFold structure of NSP8 was used for the visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP8 sequence dataset were GLN24ARG, THR145ILE, THR141MET, THR148ILE, VAL34PHE, THR89ILE, THR123ILE, PRO178HIS, ALA14THR and GLN158HIS. The mutational and conservational threshold selected were 0.096% and 0.004% respectively (Figure 2.26. a). As per site finder in MOE, a total of 13 binding pockets (Figure 2.26. b,c,d) were identified and among them site 1 was the largest binding pocket with the highest number of conserved amino acid residues (Figure 2.26. c,d). Currently, there are no known inhibitor of NSP8, thus the conserved binding pocket can be explored for the identification of novel inhibitors.

(a)



(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	29	1.69	16	28	1:(PHE92 LEU95 ARG96 ASN104 ILE106 ILE107 ASN108 ALA110 ARG111)
2	19	1.54	7	13	1:(TYR135 TYR138 LYS139 ASN140 THR141 CYS142 ASP143 THR146 PHE147 THR148 ILE172)
3	12	1.47	11	16	1:(ILE156 GLN157 GLN158 VAL159 ILE166 VAL167 GLN168 LEU169 THR187)
4	12	0.42	5	8	1:(TYR138 CYS142 ASP143 GLY144 THR145 LEU169)
5	4	0.16	5	9	1:(GLU77 ARG80 ALA81 THR84)
6	9	-0.04	7	11	1:(ASP161 LYS165 ILE166 VAL167 GLU171 ASN179 LEU180 ALA181)
7	16	-0.25	5	15	1:(SER4 GLU5 SER7 SER8 ALA53 GLN56 ARG57 GLU60)
8	5	-0.27	5	12	1:(THR84 MET87 GLN88 LEU91)
9	16	-0.38	6	12	1:(ALA102 ILE106 ILE120 LEU122 ALA150 SER151 ALA152 ARG190 SER193)
10	4	-0.54	3	13	1:(ASN104 ASN105 ASN108 ASN109 ASP112)
11	7	-1.27	4	7	1:(ASN100 ALA102 LEU103 ILE106)
12	4	-1.27	6	8	1:(LYS37 LEU38 LYS40 SER41)
13	4	-1.27	3	3	1:(ARG57 LYS58 LYS61)

(d)

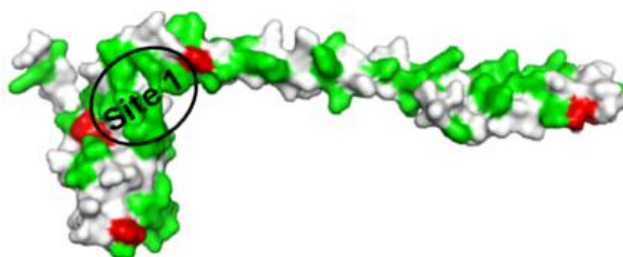


Figure 2.26. Conservational and mutational analysis of NSP8 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding site 1.

2.9.15. NSP9

The PDB ID 8WKE with a loop modelled structure was used for the visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP9 sequence dataset were ARG39LYS, ILE65VAL, THR35ILE, THR24ILE, LEU42PHE, LYS36ASN, THR109ILE, PRO57SER, PRO83LEU and PRO80LEU. The mutational and conservational threshold selected were 0.127% and 0.003% respectively (Figure 2.27. a). As per site finder in MOE, a total of 6 binding pockets (Figure 2.27. b,c,d) were identified and among them site 4 had highest number of conserved amino acid residues and looked more reasonable than other binding pockets (Figure 2.27. c,d). Currently, there are no known inhibitor for NSP9, thus the conserved binding pocket can be explored for the identification of novel inhibitors.

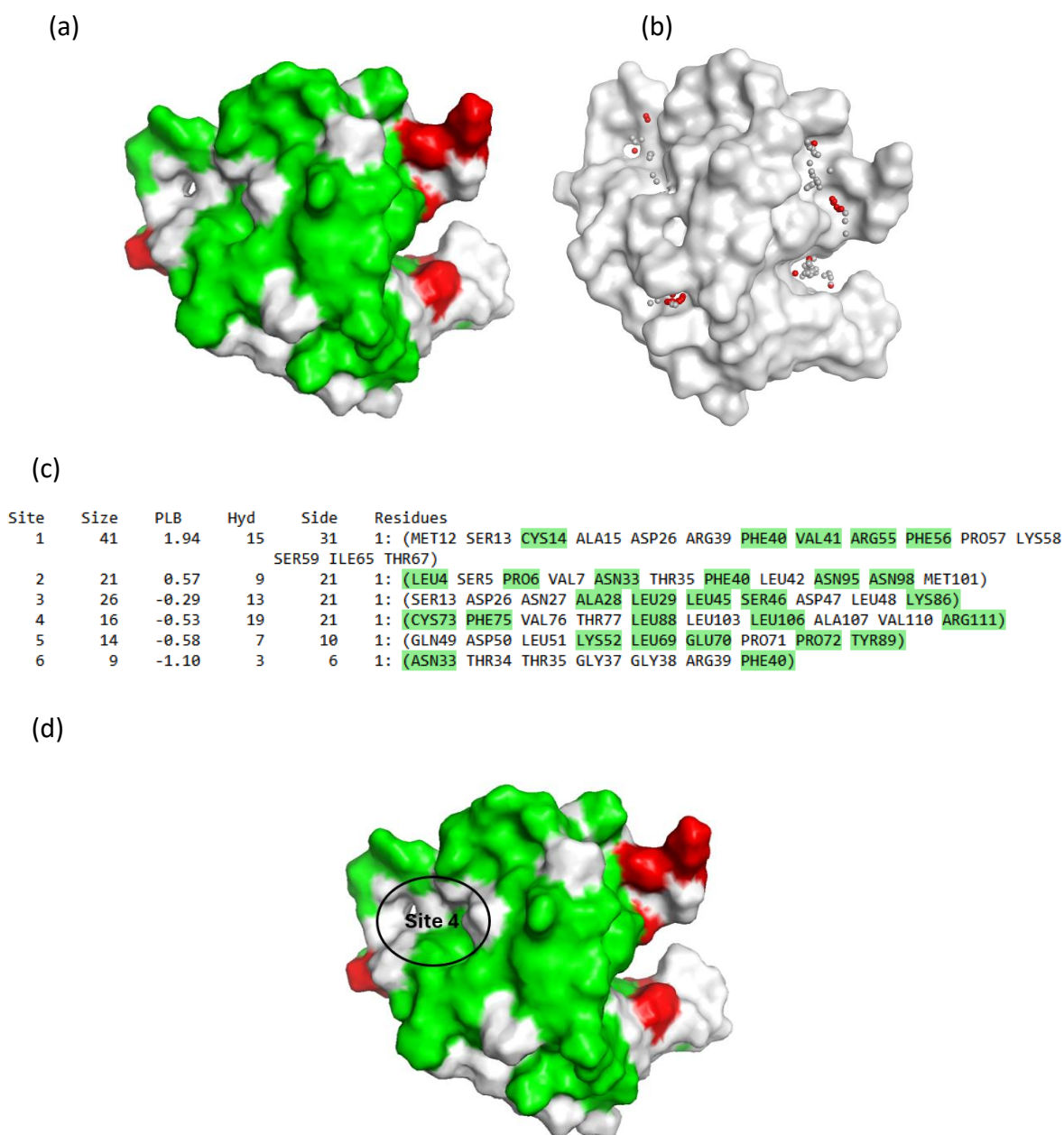
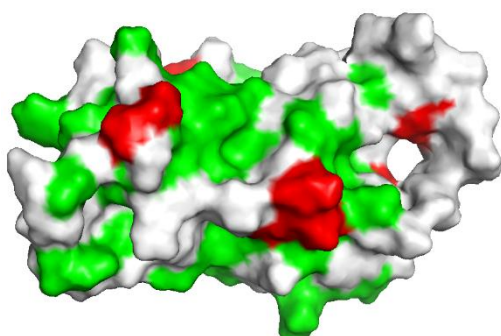


Figure 2.27. Conservation and mutational analysis of NSP9 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding site 4.

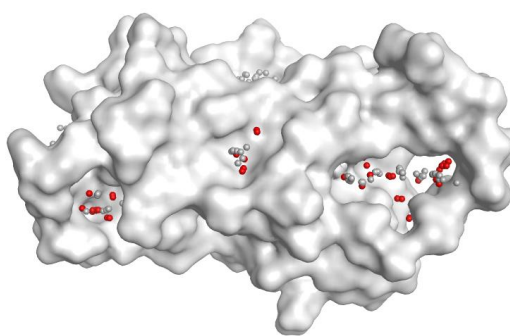
2.9.16. NSP10

The PDB ID 8VUO with a loop modelled structure was used for the visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP10 sequence dataset were THR102ILE, ALA32VAL, THR51ILE, SER33ASN, THR101ILE, ALA104VAL, ARG134CYS, PRO84SER, THR111ILE and ILE81THR. The mutational and conservational threshold selected were 0.056% and 0.003% respectively (Figure 2.28. a). As per site finder in MOE, a total of 8 binding pockets (Figure 2.28. b,c,d) were identified and among them sites 1 and 5 had the highest number of conserved amino acid residues (Figure 2.28. c,d). Currently, there are no known inhibitor of NSP10, thus the conserved binding pocket can be explored for the identification of novel inhibitors.

(a)



(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	94	1.73	25	43	1: (ILE55 CYS73 CYS74 LEU75 TYR76 HIS83 PRO84 ASN85 CYS90 ASP91 LEU92 LYS95 VAL108 THR111 LEU112 ASN114 THR115 VAL116 CYS117 THR118 GLY121 TRP123)
2	69	0.83	16	37	1: (HIS48 THR49 GLY50 THR51 GLY52 GLN53 ASN62 MET63 ASP64 GLN65 PRO100 MET122 TRP123 TYR126 GLY127 ARG134 GLU135 MET137 LEU138 GLN139)
3	31	0.76	12	29	1: (ALA1 ALA4 THR5 GLU6 VAL7 PRO8 ALA9 ASN10 SER11 THR12 VAL13 TYR30 GLN36 PRO37 ILE38 THR39)
4	29	0.35	14	21	1: (CYS17 ALA18 PHE19 ALA20 VAL21 ASP22 ALA23 ALA24 TYR76 ILE81 ASP82 PRO84)
5	13	-0.58	8	18	1: (CYS103 ASN105 ASP106 GLY109 PHE110 LYS113 ASN114)
6	13	-0.81	6	12	1: (ASP82 HIS83 GLY88 PHE89 CYS90 LYS93)
7	13	-1.09	5	9	1: (ALA9 ASN10 SER11 LEU14 SER15 ALA18 ARG78 CYS79 HIS80)
8	4	-1.18	8	9	1: (ALA1 THR12 SER15 PHE16 PHE19)

(d)

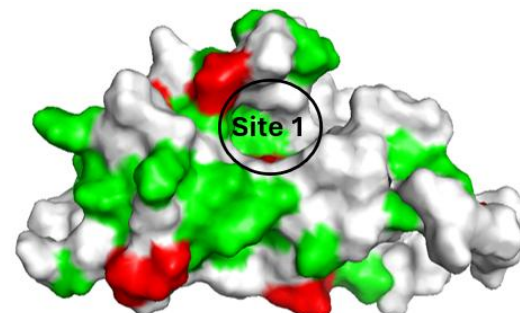
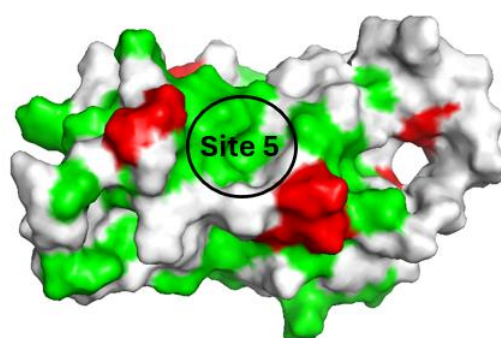
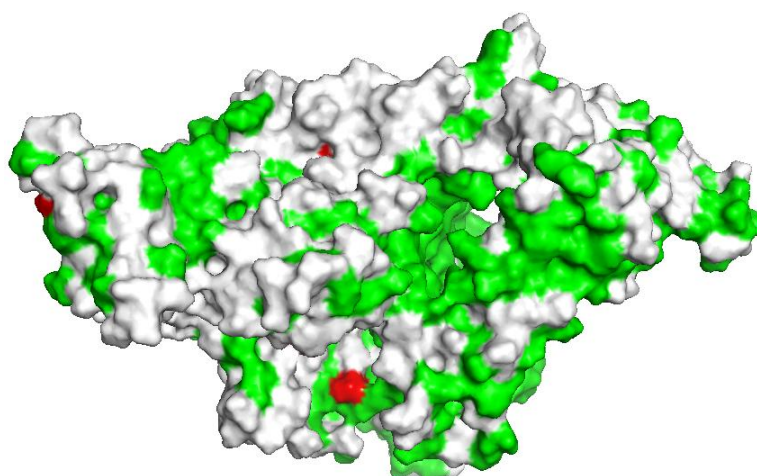


Figure 2.28. Conservational and mutational analysis of NSP10 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hvd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding sites 1 and 5.

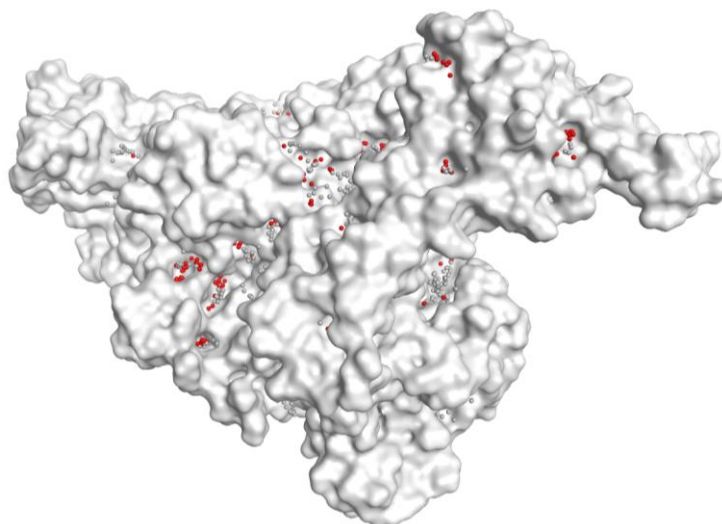
2.9.17. NSP12

The PDB ID 9DJ8 with fully resolved residues was used for the visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP12 sequence dataset were GLY671SER, PHE192VAL, PRO227LEU, PRO323LEU, ARG249MET, ALA529VAL, PHE694TYR, SER6LEU, VAL776LEU and ALA185VAL. The mutational and conservational threshold selected were 0.222% and 0.002% respectively (Figure 2.29. a). As per site finder in MOE, a total of 59 binding pockets (Figure 2.29. b,c,d) were identified and among them sites 1 and 3 had the highest number of conserved amino acid residues. The sites 13, 15, 19, 38 and 40 when joined together makes up a highly conserved binding pocket referred here as “combined site” (Figure 2.29. c,d). A known inhibitor Remdesivir (Yin et al., 2020) binds to this combined site of NSP12 (Figure 2.29. e) which is a conserved binding pocket across the SC-2 variants in our analysis, it has the potential to be used as a pan SC-2 inhibitor.

(a)



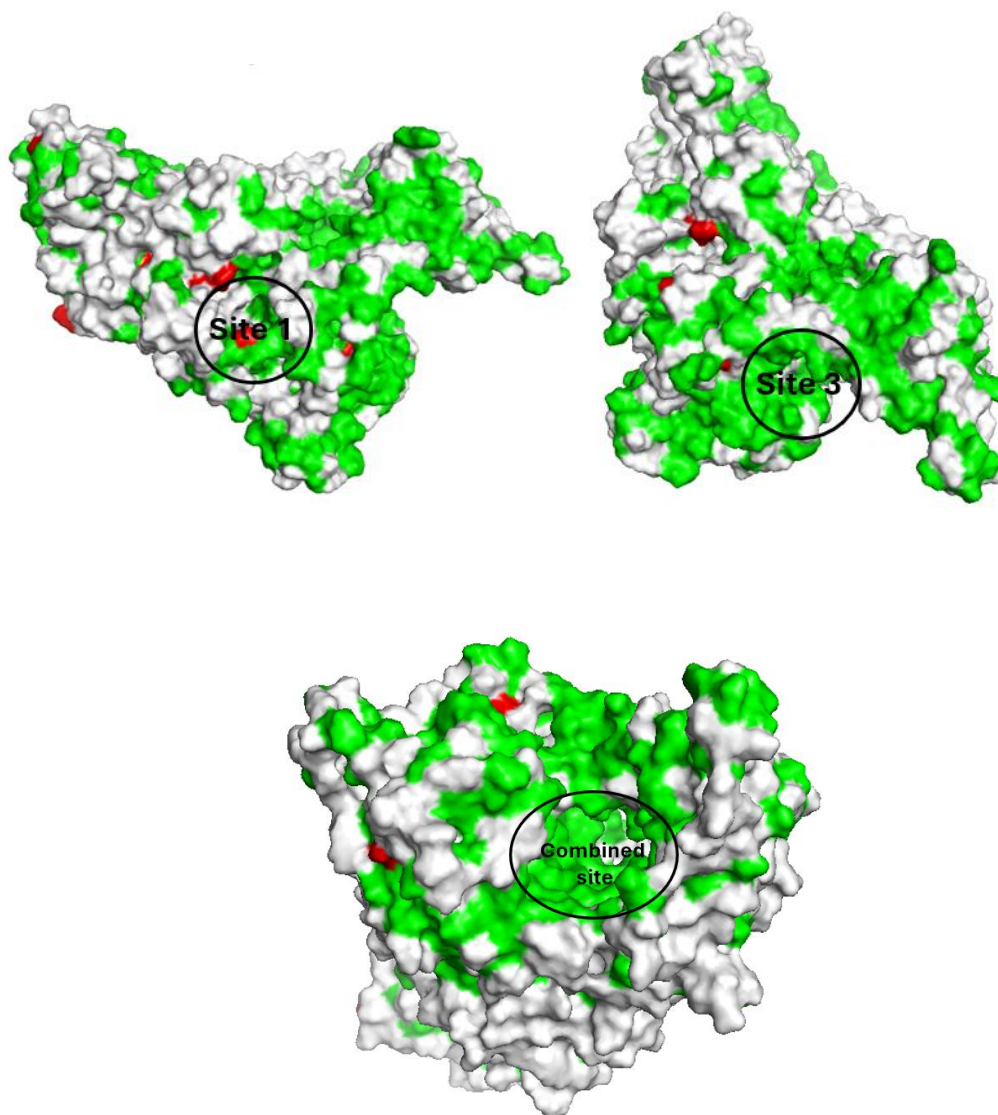
(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	97	3.79	37	57	1: (LEU270 LEU271 LYS272 TYR273 PRO323 THR324 SER325 PHE326 GLY327 PRO328 LEU329 VAL330 TYR346 HIS347 PHE348 ARG349 GLU350 PHE396 VAL398 ASN628 LEU663 SER664 GLU665 MET666 VAL675 LYS676 PRO677)
2	41	2.26	16	25	1: (VAL410 PHE440 PHE441 PHE442 GLN444 LEU544 LYS545 TYR546 ALA547 ILE548 SER549 LYS551 ASN552 ARG553 ALA554 ARG555)
3	76	2.25	32	49	1: (ILE494 VAL495 ASN496 ASN497 LEU498 LYS500 SER501 ALA502 ALA512 TYR516 GLY559 VAL560 SER561 ILE562 THR565 ARG569 GLN573 LEU576 LYS577 ALA580 GLY683 ASP684 ALA685 TYR689)
4	87	2.21	27	62	1: (VAL31 ARG33 ALA34 PHE35 ILE37 LYS50 CYS53 ARG55 VAL71 LYS73 ARG116 LEU119 THR120 LYS121 TYR122 THR123 ALA125 ASP126 ASP208 ASN209 ASP211 TYR217 ASP218)
5	60	2.04	29	51	1: (ILE333 VAL335 VAL342 SER343 THR344 ASN356 ASP358 VAL359 ASN360 SER363 SER364 ARG365 LEU366 LYS369 GLU370 VAL373 TYR374 ASP377 TYR530 ASN534)
6	20	1.25	16	29	1: (LEU388 ASP390 ARG392 VAL405 ALA406 PHE407 ALA449 ILE450 CYS669 GLY670 GLY671 SER672 TYR674)
7	84	1.24	20	39	1: (VAL315 SER318 THR319 VAL320 PHE321 PRO322 PRO323 PHE326 ARG349 GLU350 ASN459 LEU460 PRO461 THR462 PRO627 ASN628 MET629 PRO677 GLY678)
8	53	1.18	22	35	1: (ARG305 ILE466 ARG467 LEU470 PHE471 GLU474 CYS730 LEU731 ASN734 ARG735 ASP736 VAL737 ASP738 PHE741 VAL742)
9	63	1.17	14	46	1: (VAL166 GLU167 ASP452 TYR455 TYR458 ARG553 ARG555 THR556 ASP618 PRO620 LYS621 CYS622 ASP623 ARG624 LYS798)
10	44	1.02	23	37	1: (ASP303 ASP304 ILE307 PHE571 HIS572 SER635 LEU636 ALA639 ARG640 LEU648 ARG651 PHE652 ARG654 LEU655 GLU658)
11	44	0.88	14	22	1: (LYS47 TYR129 ALA130 HIS133 PHE134 ASP135 ASN138 CYS139 ASP140 THR141 LEU142)
12	52	0.80	19	38	1: (ASP36 ILE37 TYR38 ASN39 VAL204 THR206 ASN209 ASP218 PHE219 GLY220 ASP221 PHE222 SER236 TYR728 ARG733)
13	19	0.62	11	16	1: (ARG392 ALA406 PHE407 GLN408 GLY446 ASN447 ALA448 ALA449 ILE450 LEU544 CYS669 GLY670)
14	24	0.40	15	28	1: (ASP67 TYR69 THR120 LYS121 TYR122 GLU144 THR148 TYR149 LEU212)
15	11	0.35	14	20	1: (ALA490 ASN491 GLN492 VAL493 ILE494 VAL495 LEU498 TYR516 TYR521 GLN524)
16	32	0.32	6	24	1: (ASP452 TYR456 MET542 THR556 VAL557 ALA558 ASP623 ARG624 LYS676 THR680 SER681 SER682)
17	37	0.25	11	25	1: (GLY616 TRP617 ASP618 ASP760 ASP761 ALA762 VAL763 LYS798 CYS799 TRP800 GLU811 PHE812)
18	27	0.13	10	15	1: (GLU254 SER255 TYR265 ILE266 TRP268 VAL320 PHE321 PRO322)
19	36	0.11	7	22	1: (TYR479 ALA585 THR586 THR604 VAL605 SER607 LEU749 ARG750 PHE753 SER754 MET755)
20	38	0.07	16	33	1: (VAL231 PRO232 VAL233 VAL234 ASP235 TYR289 TRP290 ASP291 ARG467 TYR732 ARG733 ASN734 ARG735)
21	35	-0.02	10	30	1: (GLN292 ASP303 ARG305 CYS306 HIS309 ARG467 LEU470 ARG735 ASP736 VAL737)
22	35	-0.04	10	30	1: (ARG467 LEU470 ARG735 ASP736 VAL737)
23	18	-0.47	6	10	1: (PHE165 VAL166 PRO169 LEU172 ARG457 TYR458 LEU460 ASN791 VAL792)
24	24	-0.47	13	20	1: (TYR32 ARG33 ALA34 LYS47 TYR122 ALA125 ASP126 TYR129 ALA130 THR141)
25	19	-0.50	9	18	1: (ASP164 VAL166 GLU167 PRO620 LYS621 PHE793 SER795)
26	9	-0.53	6	10	1: (PHE429 LYS430 GLU431 GLY432 SER433 VAL435 GLU436 LEU437)
27	13	-0.58	16	16	1: (GLY413 ASN414 PHE415 PHE440 PHE441 PHE443)
28	12	-0.59	9	17	1: (ASP161 TYR163 ASP164 GLU167 PHE793 SER795)
29	13	-0.62	5	13	1: (ALA176 ASN177 GLY179 GLU180 THR248 ARG249 ALA250 THR252 LEU261)
30	8	-0.65	4	13	1: (ASP618 PRO620 LYS621 PHE793 MET794 SER795 LYS798)
31	19	-0.70	6	17	1: (GLN444 ALA448 SER451 ASP452 TYR455 ARG553 ALA554)
32	28	-0.73	12	17	1: (LEU172 ARG173 ALA176 THR246 LEU247 ARG249 LEU460 PRO461)
33	10	-0.74	5	11	1: (ARG349 CYS395 PHE396 VAL675 LYS676 PRO677)
34	13	-0.77	9	18	1: (GLU474 ASP477 LYS478 VAL737 THR739 VAL742 ASN743)
35	16	-0.77	2	9	1: (GLU802 LYS807 GLY808 PRO809 HIS810 HIS816 THR817 TYR831)
36	24	-0.82	10	15	1: (PRO169 LEU172 THR394 CYS395 TYR456 ARG457 TYR458 ASN459 PRO677)
37	10	-0.86	11	16	1: (ILE856 GLU857 VAL860 HIS892 TYR915 PHE920)
38	8	-0.95	7	7	1: (LEU187 VAL190 GLN191 ARG285)
39	14	-1.01	2	7	1: (GLN57 LEU65 ILE66 ASP67 TYR69 GLU144 THR148)
40	15	-1.01	4	9	1: (TYR32 LYS47 TYR129)
41	18	-1.01	2	3	1: (LYS500 SER501)
42	14	-1.01	3	13	1: (ASP303 ASP304 ARG640 THR643 ARG651)
43	16	-1.01	6	18	1: (HIS892 ASP893 THR896 THR908 ASN909 THR912 TRP916)
44	15	-1.01	15	17	1: (PHE368 LEU371 LEU372 ALA375 PHE506 TRP509 LEU514 TYR515)
45	5	-1.01	3	11	1: (ASP67 THR148 TYR149 ASN150)

(d)



(e)

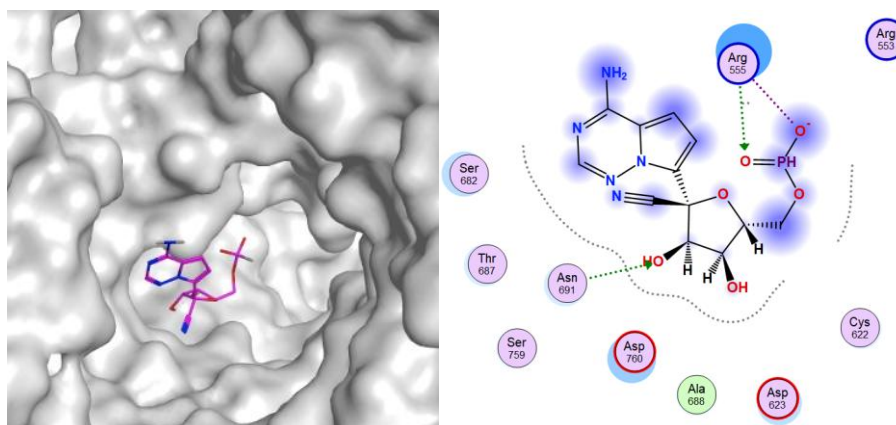


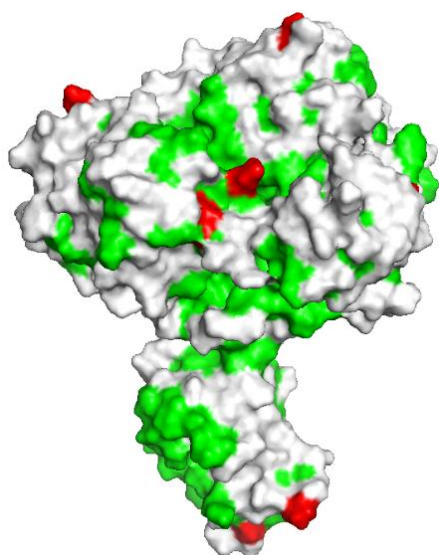
Figure 2.29. Conservational and mutational analysis of NSP12 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with

hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding sites 1 and 3 as well as the “combined site”. (e) Remdesivir (pink) inhibitor within binding pocket (grey) (PDB ID 7BV2) with its protein interaction map. Hydrogen bonds are depicted in dotted lines.

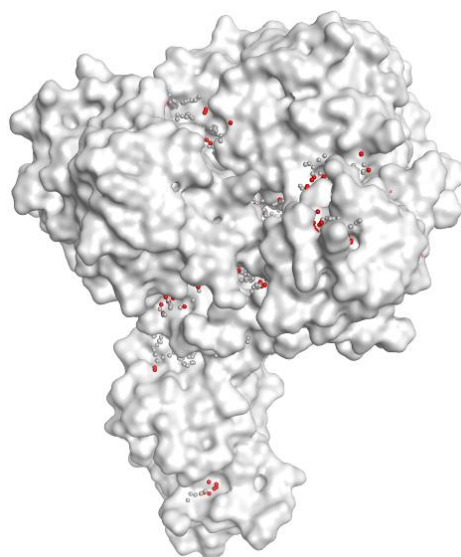
2.9.18. NSP13

The PDB ID 5RML with fully resolved residues was used for the visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP13 sequence dataset were PRO77LEU, ILE334VAL, HIS164TYR, LYS460ARG, ALA296SER, GLN88HIS, ASP260TYR, GLU341ASP, PRO53LEU and PRO504LEU. The mutational and conservational threshold selected were 0.437% and 0.002% respectively (Figure 2.30 a). As per site finder in MOE, a total of 25 binding pockets (Figure 2.30. b,c,d) were identified and among them site 1 had the highest number of conserved amino acid residues (Figure 2.30. c,d). There are fragments co-crystalised at different binding pockets of NSP13 (Newman et al., 2021) which are conserved in our analysis (discussed in Chapter III). A very recent study by Herasymenko et al., (2025) identified scaffolds binding to NSP13 with a K_d below 10 μ M and inhibited *in-vitro* helicase activity. These scaffolds targeted the RNA-binding tunnel of NSP13 which was found to be highly conserved (site 1) in our SC-2 variants analysis, they have the potential to be used as a pan SC-2 inhibitors.

(a)



(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	202	4.23	57	124	1: (LYS139 GLU142 GLU143 PHE145 LYS146 TYR149 ASN179 TYR180 VAL181 GLU197 LEU227 THR228 SER229 HIS230 MET233 PRO234 LEU235 SER236 ALA237 ALA308 CYS309 SER310 HIS311 PRO335 ARG337 ARG339 THR359 VAL360 ASN361 ALA362 GLU365 THR380 TYR382 ASP383 SER385 VAL386 ALA389 ARG390 ALA407 PRO408 ARG409 THR410 LEU411) 2: (TYR120 ALA123 ASN124 ALA135 LEU138 LYS139 GLU142 THR380 ASN381 TYR382 ASP383 ARG409 LEU417 TYR421 PHE422) 3: (GLU261 SER264 ASN265 GLY285 THR286 GLY287 LYS288 SER289 HIS290 PHE291 ALA316 GLU319 LYS320 LEU438 GLY439 THR440 ARG442 ARG443 GLU540) 4: (ARG155 LEU163 HIS164 LEU165 TRP167 LEU176 LYS202 GLY203 ASP204 TYR205 GLY206 ASP207 ALA208 VAL209 ASP483) 5: (ASN177 ARG178 ASN179 VAL181 GLU197 TYR198 THR199 ARG212 THR214 CYS309 SER310 HIS311 ALA312 ASP315 ARG339 VAL340 GLU341 MET378 ASP383 LEU405 PRO406 ALA407 PRO408 ASN516 THR532 ASP534 SER535) 6: (GLU162 ARG178 GLU201 LYS202 ASP204 VAL210 ARG212 ASN516 ASN519 ALA520 VAL521 SER523 LYS524 THR530) 7: (ASN46 PRO47 VAL49 ASN51 MET68 SER69 TYR70) 8: (MET474 PHE475 TYR476 ARG579 TYR582 SER589 LEU590 GLU591 ILE592 PRO593 ARG594) 9: (VAL157 SER159 ASP160 TYR198 TYR211 ARG212 GLY213 THR215 THR216 TYR217 LYS218 LEU219) 10: (PRO283 PRO284 GLY285 THR286 GLY287 LYS288 SER289 HIS290 SER310 ALA312 ALA313 ALA316 LEU317 ASP374 GLU375 GLN404 ARG443 GLN537 GLY538 ARG567) 11: (GLY3 ALA4 CYS5 VAL6 ARG21 ARG22 PRO23 PHE24 ARG129 LEU132 PHE133 GLU136 PRO234 LEU235 SER236) 12: (PRO406 ALA407 PRO408 ARG409 THR410 LEU412 THR413 LYS414 GLY415 THR416 LEU417 PHE422 ALA553 HIS554 ASN557 ASN559 ARG560) 13: (ILE20 ILE109 ALA110 THR111 CYS112 ASP113 TRP114 THR137 THR141) 14: (GLN281 PRO402 ALA403 MET429 LYS430 GLY433 PRO434 PHE437 VAL456 TYR457 ASP458 LYS460) 15: (TRP167 ARG173 PRO174 PRO175 LEU176 ASN177 LYS202 ASP483 VAL484 SER486 ASN516) 16: (PHE81 PRO82 LEU83 CYS84 ALA85 ASN86 GLN88 VAL89 PHE90 LYS94) 17: (LYS202 GLY203 ASP204 THR481 HIS482 ASP483 ILE488 ARG490 SER517 ALA520 VAL521) 18: (TYR149 PRO175 ASN177 ASN179 TYR180 SER485 SER486 PRO514 TYR515 ASN516 HIS554) 19: (ARG21 ALA140 GLU143 THR144 THR231 VAL232) 20: (LYS146 LEU147 SER148 TYR149 GLY150 ILE151 PRO174 PHE182 PHE225 VAL226 LEU227 SER229) 21: (GLU128 LYS131 THR239 LEU240 SER424 LEU428) 22: (SER13 LEU14 SER44 VAL45 ASN46 TYR48 TYR70 PHE90 GLY91 LEU92) 23: (SER10 CYS26 CYS27 LYS28 GLN88 LYS94) 24: (GLY538 SER539 GLU540) 25: (THR183 GLN194 GLY196 GLU197 THR228 ARG337)

(d)

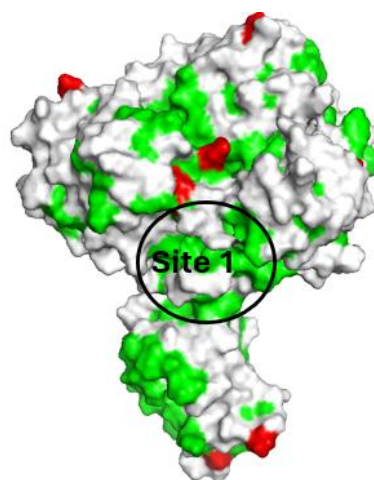


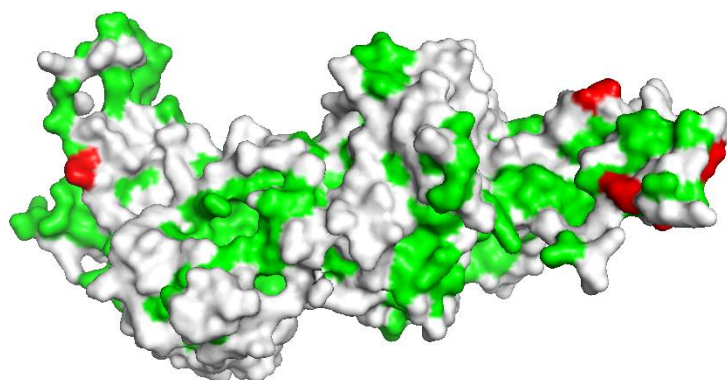
Figure 2.30. Conservational and mutational analysis of NSP13 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity

(Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d)
Visualisation of conserved binding site 1.

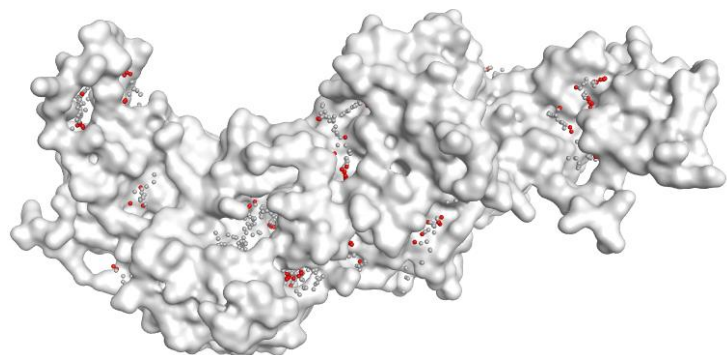
2.9.19. NSP14

The AlphaFold predicted structure was used for visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP14 sequence dataset were ALA394VAL, ILE42VAL, ASN129ASP, ALA323SER, PRO46LEU, PRO451SER, SER374PHE, HIS26TYR, THR31ILE and MET72ILE. The mutational and conservational threshold selected were 0.376% and 0.003% respectively (Figure 2.31. a). As per site finder in MOE, a total of 28 binding pockets (Figure 2.31. b,c,d) were identified and among them sites 1 and 2 had the highest number of conserved amino acid residues (Figure 2.31.. c,d). A study by Meyer et al., 2025 identified a NSP14 inhibitor, TDI-015051 (Figure 2.31.e) which occupies the highly conserved sites 1 and 2 in our SC-2 variants analysis, hence it might have the potential to be used as a pan SC-2 inhibitor.

(a)



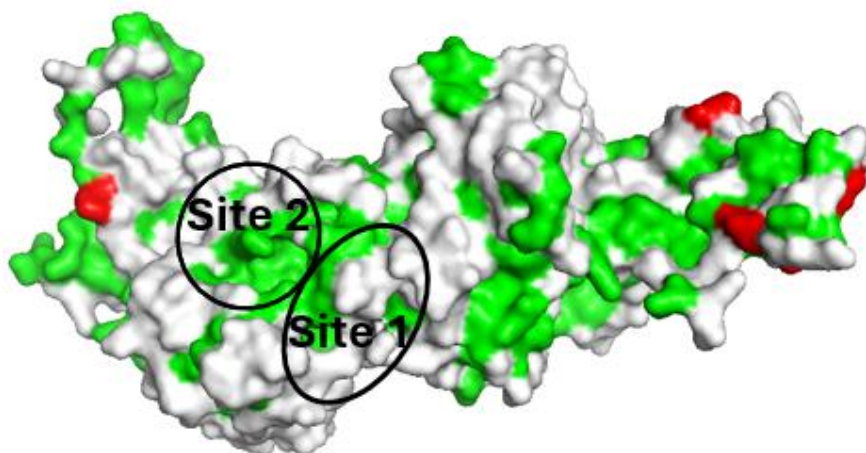
(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	191	4.14	48	80	1: (PHE286 VAL287 LYS288 ARG289 VAL290 ASP291 TRP292 THR293 ILE332 GLY333 PRO335 TYR351 ASP352 ALA353 GLN354 PRO355 CYS356 LEU366 PHE367 TYR368 SER369 TRP385 ASN386 CYS387 ASN388 VAL389 ASP390 CYS399 ARG400 PHE401 TYR420 PHE426 HIS427 THR428 PRO429 PHE506) 1: (TRP292 ILE305 ASN306 CYS309 ARG310 GLN313 ASP331 ILE332 GLY333 ASN334 PRO335 LYS336 ILE338 LYS339 CYS340 LEU383 PHE384 TRP385 ASN386 TYR420 ASN422 LYS423 HIS424 PHE426 PHE506)
2	106	2.14	37	74	1: (ARG76 ILE80 GLU284 CYS285 PHE286 LYS288 ARG289 VAL290 ASP291 ILE294 TYR296 HIS424 ALA425)
3	43	1.21	22	34	1: (ASP90 VAL91 GLU92 GLY93 CYS94 HIS95 ASN104 PRO141 PHE146 PHE190 HIS268 ASP273) 1: (ILE80 ARG81 HIS82 VAL83 ARG84 LYS175 ASN176 LEU177 SER178 ASP179 ARG212 HIS228 TYR296 PRO297 ILE299)
4	30	0.59	10	26	1: (LEU174 LYS175 SER178 ASP179 ARG212 ARG213 THR215 CYS226 TRP227 HIS228)
5	54	0.32	11	32	1: (VAL83 ARG84 ALA85 TRP86 ILE87 SER112 THR113 THR173 LEU174 LEU177 ARG278 LEU411 PRO412)
6	14	0.06	8	20	1: (GLU99 VAL120 PRO121 ARG135 VAL136 SER137 ALA138 MET153)
7	30	0.01	11	19	1: (LEU18 THR21 GLN22 PRO24 TYR51 ASP126 THR127 PRO128 ASN129 ASN130 THR131) 1: (VAL120 THR122 PHE133 SER134 ARG135 LEU157 PRO158 TRP159 ASN160 ARG163) 1: (VAL167 GLN168 SER171 ASP172 LYS175 TRP227 HIS228 HIS229 SER230 ILE231 GLY232) 1: (ALA119 VAL120 PRO121 ARG135 LEU152 MET153 TYR154 LYS155 GLY156) 1: (LEU321 LEU322 LYS325 PHE326 ALA344 ASP345 VAL346) 1: (MET62 ASN63 TYR64 GLN65 ASN71 MET72 PHE73) 1: (PHE73 ILE74 THR75 ARG76 GLU77 ILE242 ASP243 GLN246 TRP247) 1: (GLU453 SER454 HIS455 GLY456 LYS457 GLN458 ASP462 TYR465 CYS484 ARG485 HIS486) 1: (THR5 LEU7 PHE8 LYS9 ASP10 PRO24 THR25 LEU27 SER28 ILE55 MET57 ASP126 THR131)
8	34	-0.01	11	22	1: (PHE89 ASP90 VAL91 LEU185 TRP186 ALA187 PHE190 GLU191)
9	25	-0.02	9	18	1: (VAL389 ASP390 ARG400 PRO429 ALA430 ASP432 SER434 ALA435)
10	22	-0.12	13	22	1: (THR173 LEU177 LEU406 SER407 ASN408 ASN410 LEU411 PRO412)
11	24	-0.21	8	17	1: (SER112 HIS264 ALA274 ILE275 ARG278 CYS279 PRO412 GLY413)
12	10	-0.21	4	9	1: (SER461 ASP462 ILE463 ASP464 TYR465 ARG485 ASN489 ARG492)
13	24	-0.21	15	17	1: (SER12 ARG52 ARG53 LEU54 GLU99 ALA100 VAL125 ARG135 VAL136 SER137)
14	10	-0.30	6	11	1: (ARG76 TRP247 PHE249 GLU284)
15	34	-0.34	9	19	1: (PHE60 MET62 MET72 LEU192 MET195 LYS196 VAL199 LYS200 ILE201)
16	46	-0.37	11	28	1: (LYS311 HIS314 MET315 LYS318 TYR446 TYR447 SER448 ASP449 SER450 GLU490)
17	40	-0.40	16	25	1: (ASP464 TYR465 VAL466 LEU468 ALA488 ASN489 ARG492) 1: (LYS469 SER470 ALA471 ASN478 GLY481 TRP520)

(d)



(e)

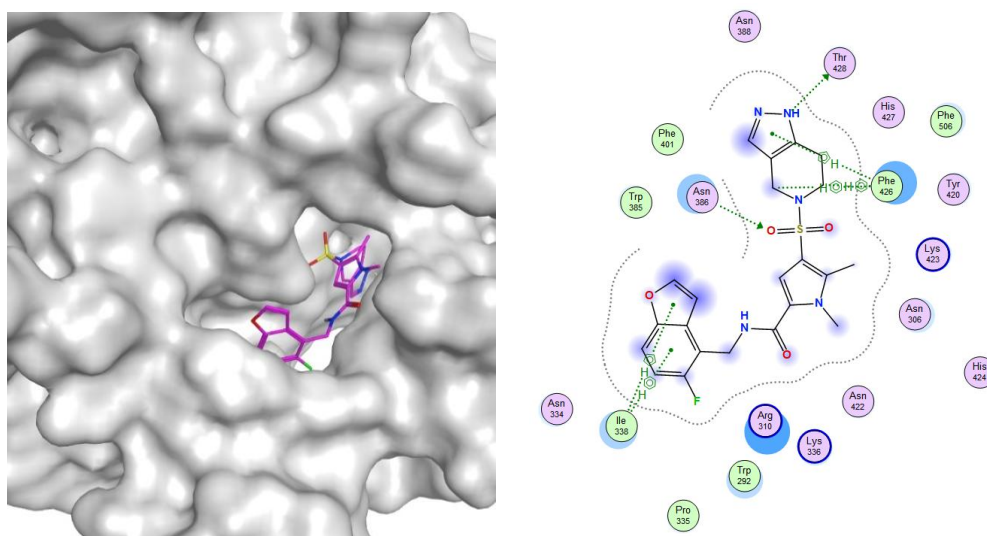


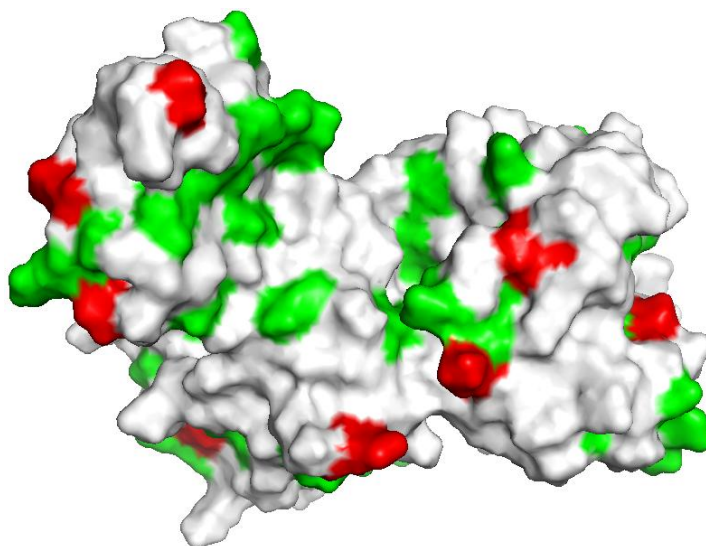
Figure 2.31. Conservational and mutational analysis of NSP14 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding sites 1 and 2. (e) TDI-015051 (pink) inhibitor within binding pocket (grey) (PDB ID 8R7B) with its protein interaction map. Hydrogen/hydrophobic interactions are depicted in dotted lines.

2.9.20. NSP15

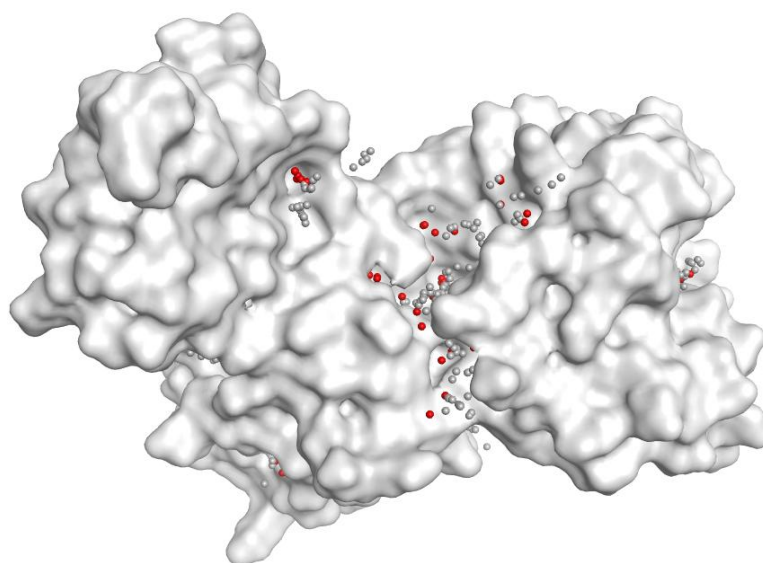
The PDB ID 7TQV with fully resolved residues was used for the visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP15 sequence dataset were LYS259ARG, ALA80VAL, ASP219TYR, ALA171SER, GLN19HIS, PRO205SER, ALA255VAL, VAL22LEU, ASP132TYR and ASP39TYR. The mutational and conservational threshold selected were 0.221% and 0.004% respectively (Figure 2.32. a). As per site finder in MOE, a total of 22 binding pockets (Figure 2.32. b,c,d) were identified and among them sites 1 and 6 had the highest number of conserved amino acid residues (Figure 2.32.c,d). There are fragments co-crystallised at different binding pockets of NSP15 (Godoy et al., 2023) which are conserved in our analysis. A study by Kim et al., (2021) identified Tipiracil (Figure 2.32.

e), an FDA approved drug that is used in the treatment of colorectal cancer, as a potential anti-COVID-19 drug. It interacts with the uridine binding pocket which was found to be highly conserved (site 6) in our SC-2 variants analysis and might have the potential to be used as a pan SC-2 inhibitor.

(a)



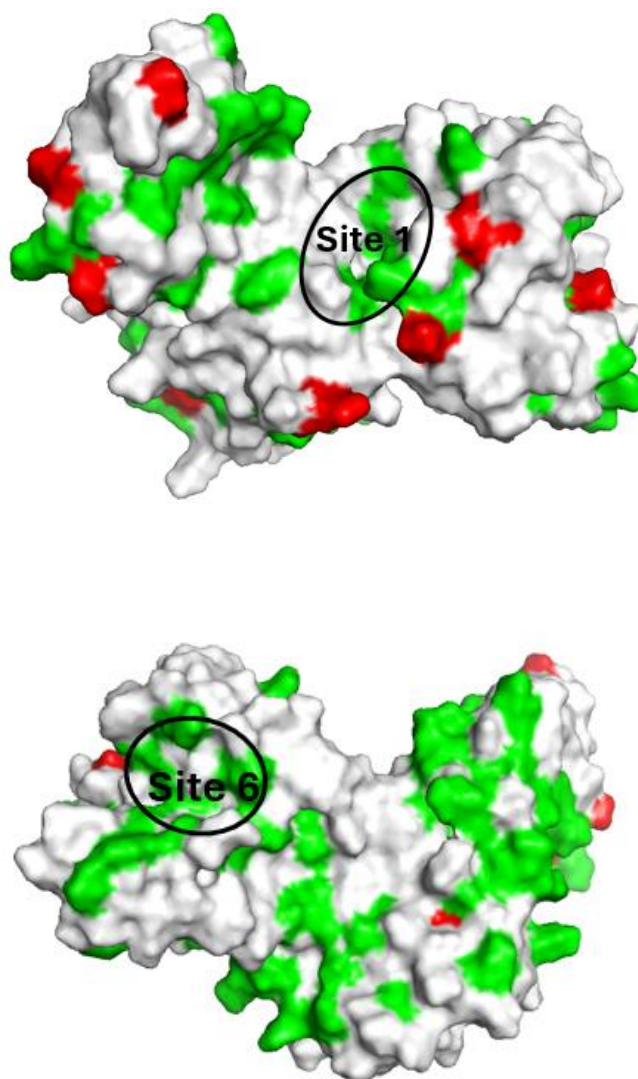
(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	143	4.16	47	79	1: (ASN63 VAL67 PRO68 GLU69 VAL70 LYS71 ILE86 TRP87 ASP88 TYR89 LYS90 SER162 GLY165 VAL166 THR167 LEU168 ILE169 THR196 GLN197 SER198 ARG199 ASN200 GLU203 PHE204 LYS205 ARG207 LEU252 ASP273 SER274 THR275 VAL276 LYS277 TYR279 VAL295 ILE296 ASP297)
2	36	0.64	13	18	1: (ASP220 ILE223 GLU224 LEU228 GLU229 GLY230 TYR231 ALA232 PHE233 GLU234 HIS338 VAL339)
3	46	0.58	23	38	1: (HIS15 PHE16 ASP17 GLY18 LYS61 ARG62 ASN63 ILE64 LYS65 ALA82 ASN83 THR84 VAL85 GLY101 VAL102 ASP125 VAL128 GLN131)
4	24	0.55	13	18	1: (ILE80 ALA81 THR84 VAL85 ILE86 PRO94 ALA95 HIS96 SER98 THR99)
5	40	0.51	23	36	1: (LYS90 ARG199 ASN200 LEU201 GLN202 LEU252 PHE259 PHE264 GLU265 LEU266 ASP268 ILE270 PRO271 MET272 ASP273 SER274 LYS277 TYR279)
6	56	0.13	17	32	1: (ALA235 ILE236 GLY247 GLY248 HIS250 ASN278 LYS290 VAL292 CYS293 SER294 GLU340 THR341 PHE342 TYR343 LYS345 LEU346 GLN347)
7	32	0.01	14	20	1: (GLN153 PRO154 SER155 VAL156 GLN189 LEU190 PRO191 GLU192 THR193 PHE195 TYR325)
8	24	-0.16	9	18	1: (PHE44 GLU45 ASN46 LYS47 THR48 TRP59 TRP87 ASP92)
9	29	-0.18	14	25	1: (MET219 GLU234 ALA235 TYR238 GLY239 ASP240 PHE241 GLY254 LYS257 ARG258 GLU261)
10	25	-0.19	11	19	1: (VAL156 GLY157 PRO158 LYS159 GLN160 GLN176 THR193 TYR194 PHE195 THR196 GLN197 LEU299 ASP302)
11	10	-0.20	10	16	1: (TYR194 GLU305 ILE306 SER309 GLN310 VAL319 VAL321)
12	19	-0.28	11	23	1: (ASP268 PHE269 ILE270 MET272 SER274 VAL276 LYS277 ASN278 LYS345)
13	23	-0.48	7	12	1: (VAL276 GLU327 ILE328 SER329 PRO344 LYS345 LEU346 GLN347)
14	9	-0.49	3	5	1: (GLN202 PRO263 PHE264 GLU265 LEU266 GLU267)
15	11	-0.50	3	10	1: (GLU57 LYS61 GLY101 VAL102 CYS103 SER104 ASP107 LYS110)
16	9	-0.53	5	8	1: (TYR194 SER309 LYS320 VAL321 THR322)
17	8	-0.54	5	10	1: (GLY126 ARG127 THR145 GLU146 GLY147)
18	10	-0.54	4	4	1: (PRO112 THR113 GLU114 THR115 CYS117 ALA118 ASN137 ALA138 ARG139)
19	11	-0.59	9	14	1: (THR48 THR49 LEU50 ARG91 ALA93 PRO94 HIS96 ILE97)
20	8	-0.60	5	14	1: (GLU146 GLY147 SER148 VAL156 LYS174 GLN176 PHE177 ASN178)
21	14	-0.63	9	10	1: (GLY151 LEU152 GLN153 TYR180 VAL187 GLN188 GLN189)
22	5	-0.67	5	7	1: (LYS47 ARG91 ASP92 PRO271)

(d)



(e)

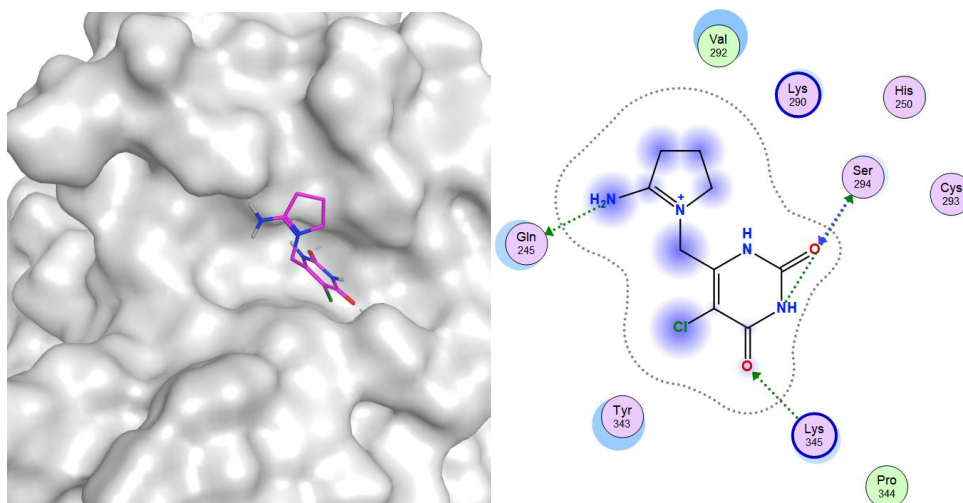
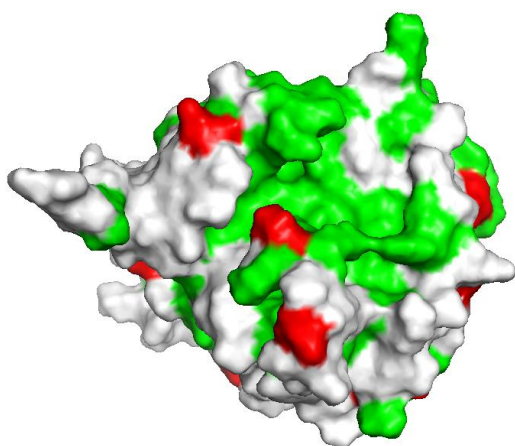


Figure 2.32. Conservational and mutational analysis of NSP15 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hvd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding sites 1 and 6. (e) Tipiracil (pink) inhibitor within binding pocket (grey) (PDB ID 6WXC) with its protein interaction map. Hydrogen bonds are depicted in dotted lines.

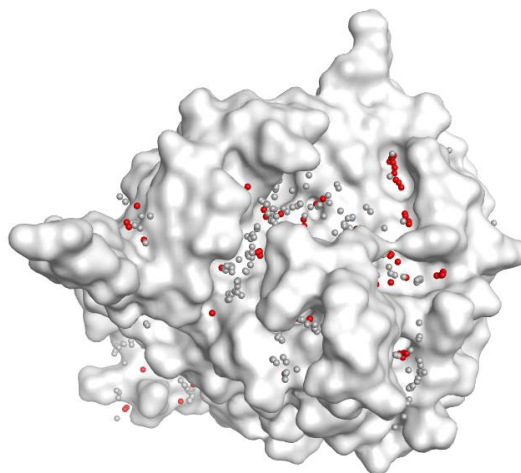
2.9.21. NSP16

The PDB ID 7JYY with fully resolved residues was used for the visualisation of the conserved/mutated residues. The top 10 mutations in the SC-2 NSP16 sequence dataset were ARG216CYS, GLN238HIS, LYS160ARG, MET20ILE, PRO215LEU, ASP102ASN, PRO134SER, PRO236SER, ASP269GLU and THR140ILE. The mutational and conservational threshold selected were 0.205% and 0.003% respectively (Figure 2.33. a). As per site finder in MOE, a total of 19 binding pockets (Figure 2.33. b,c,d) were identified and among them site 1 had the highest number of conserved amino acid residues (Figure 2.33.c,d). A study by Kalnins et al., (2024) identified multiple S-adenosyl methionine (SAM) derived inhibitors targeting the binding site of NSP16. An example of inhibitor 1 MTA is shown in (Figure 2.33. e) which binds to a highly conserved binding site (site 1) in our SC-2 variants analysis and might have the potential to be used as a pan SC-2 inhibitor.

(a)



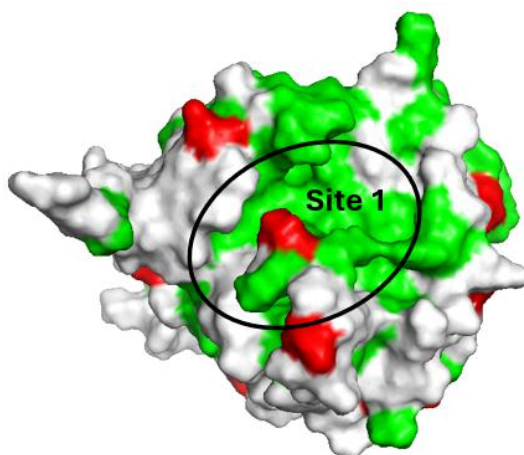
(b)



(c)

Site	Size	PLB	Hyd	Side	Residues
1	248	4.03	63	141	1: (CYS25 ASP26 LEU27 TYR30 GLY31 ASP32 ASN43 LYS46 PHE70 GLY71 ALA72 GLY73 SER74 ASP75 LYS76 SER98 ASP99 LEU100 ASN101 GLY113 ASP114 CYS115 ALA116 ASP130 MET131 TYR132 ASP133 PRO134 LYS135 THR136 LYS137 ASN138 VAL139 THR140 LYS141 GLU142 ASP144 SER145 LYS146 GLU147 GLY148 PHE149 PHE150 LYS170 THR172 GLU173 HIS174 SER175 TRP176 ASN198 SER200 SER201 SER202 GLU203)
2	71	0.70	24	43	1: (PRO7 GLN52 LEU54 ASN55 THR56 LEU57 THR58 LEU59 VAL61 SER243 ASP246 MET247 SER248 LYS249 PHE250 PRO251 LEU252 LYS253 ARG255)
3	64	0.25	22	41	1: (ARG66 ALA121 ASN122 LYS123 GLN158 LEU161 ALA162 LEU163 TYR211 GLY213 SER261 LEU262 LYS263 GLU264 ILE267 ILE282 ARG283 GLU284 ASN285 ASN286)
4	41	0.21	18	42	1: (ASN13 LYS16 MET17 GLN18 ARG19 TRP189 TRP190 ARG216 GLU217 GLN218 ILE219 ASP220 VAL223 MET224 ASN227)
5	18	0.19	15	22	1: (GLN3 GLN6 PRO7 GLY8 VAL9 ALA10 TYR53 THR56 PHE193)
6	18	-0.26	7	19	1: (ASN43 TYR47 HIS69 GLY71 ALA72 GLY73 SER74 ALA79 PRO80 GLY81 THR82 ASP130)
7	21	-0.27	11	18	1: (LEU57 THR58 ALA188 TRP189 CYS209 ASN210 SER276)
8	14	-0.30	9	16	1: (PRO62 TYR63 ASN64 GLY256 THR257 ALA258 VAL259 ARG283 ASN285)
9	14	-0.32	7	14	1: (ASP32 SER33 ALA34 LEU36 MET42 VAL197 ASN198 SER200 LEU239)
10	11	-0.33	5	9	1: (LEU111 ILE112 GLY113 THR117 SER292 ASP293 VAL294)
11	11	-0.34	7	15	1: (VAL118 HIS119 THR120 TYR152 PHE156 GLN159 LYS160)
12	32	-0.34	11	21	1: (LEU22 GLU23 LYS24 VAL139 LYS141 GLU142 ASN143 ASP144 GLU173 HIS174)
13	13	-0.39	4	10	1: (ALA34 THR35 LEU239 SER240 SER241 TYR242)
14	5	-0.41	4	6	1: (VAL78 ALA79 PRO80 ALA83 VAL104 SER105 ASP106)
15	3	-0.47	5	5	1: (PRO12 ASN13 LEU14)
16	18	-0.47	13	17	1: (LEU163 LEU212 GLY213 LYS214 GLU264 ILE267 LEU272)
17	6	-0.48	5	5	1: (LYS263 GLU264 GLY265)
18	19	-0.50	12	21	1: (PRO37 ILE40 VAL44 ALA45 THR48 GLN87 LEU244 MET247)
19	7	-0.50	4	9	1: (GLN28 ASN29 PRO236 ILE237 GLN238)

(d)



(e)

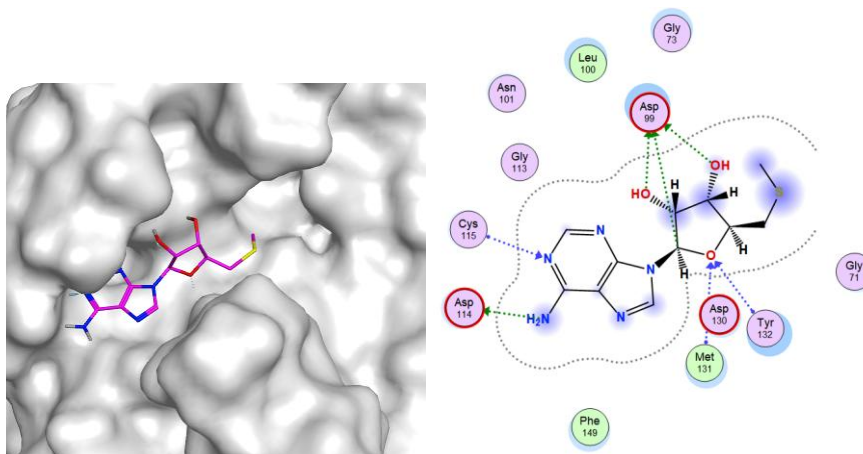


Figure 2.33. Conservational and mutational analysis of NSP16 (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues. (b) Site finder dummy atoms flooding the putative binding pockets with hydrophobic (grey) and hydrophilic (red) cavity points. (c) List of putative binding pockets with conserved amino acid residues highlighted in green. The site finder also resulted in the total size of the binding site (size), propensity for ligand binding (PLB), hydrophobicity (Hyd), and number of side-chain atoms (side). Note HIP refers to protonated histidine. (d) Visualisation of conserved binding site 1. (e) Inhibitor 1 MTA (pink) inhibitor within binding pocket (grey) (PDB ID 8RV5) with its protein interaction map. Hydrogen bonds are depicted in dotted lines.

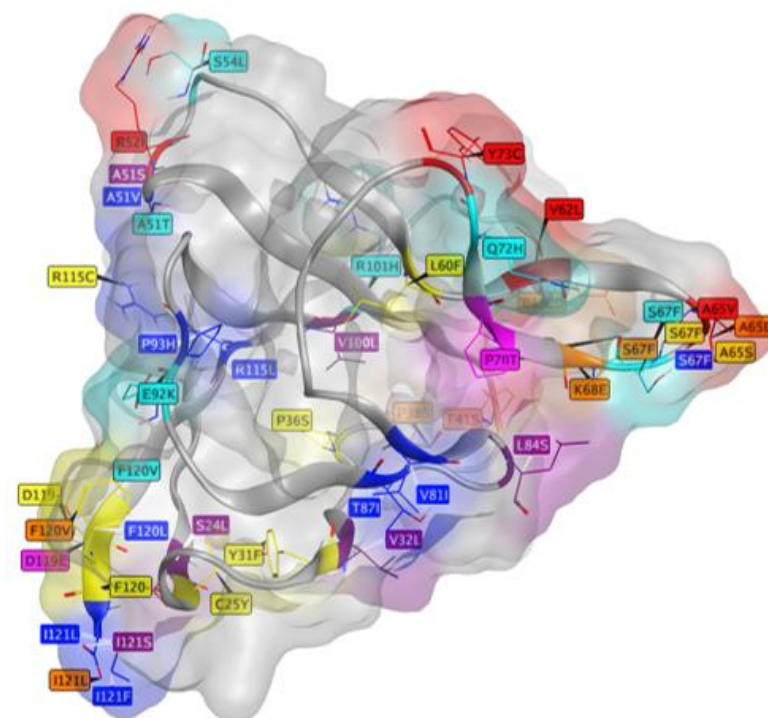
2.10. Discussion

The sequence similarity between SC-2 and SC-1 is quite striking. This discovery can be the foundation for medicinal chemistry studies for therapeutic developments for the infection caused by these viruses. The combined efforts of various research disciplines, such as bioinformatics, computational drug design, virology and clinical studies, can help speed up the task of finding therapeutics for COVID-19. Most of the drug development studies have focused on targeting the viral proteases, polymerases, and structural proteins. There is a need to explore the NSPs as potential targets for treating COVID-19. Bioinformatics tools, such as MSA studies of viral proteins of SC-2 with other related SARS-CoV viruses, can be very useful for identifying conserved residues. These residues may, in turn, define binding sites or clefts that can be initially targeted by repurposed drugs or, subsequently, tailored small molecule inhibitors. For example, the crystal structure of RdRp bound to Remdesivir

shows that it forms interactions with conserved residues K545 and R555 present in RdRp of SC-2 (Yin et al., 2020). Sinefungin, co-crystallised with NSP16, is known to interact with the conserved residues present inside the active site (Krafcikova et al., 2020). Hence, with the help of a comprehensive sequence alignment study on the SC-2 virus, the conserved regions that have a role in protein activity could be identified and targeted for further binding site analysis with suitable inhibitors. In our study NSP11 was not included as it is a very small protein (13 amino acids) which is a cleavage product processed by Mpro at nsp10/11 boundary. The first nine amino acids of NSP11 are present in NSP12. Due to its small length, it exhibits an intrinsically disordered protein behaviour (Gadhav et al., 2021).

The statistical analysis in our study showed that the NCBI dataset has proportionally more submissions from the USA than the GISAID dataset, which might lead to bias within our analysis. Thus, a new database was considered to be included in the study, i.e., GISAID, particularly for NSP3 and NSP13 (discussed in Chapter III), which has sequence submissions from around the world. The statistical analysis on the GISAID dataset indicated that the sequence distribution was better than NCBI, even though the majority of the submissions were still from the USA. The variant distribution analysis showed that in the NCBI dataset, the largest number of sequences belonged to the Delta variant. The analysis helped in understanding the spread of the sequences considering the type of SC-2 variants. As a part of the bioinformatics project, we also performed variant analysis and conservational/mutational analysis on ~1 million SC-2 ORF8 protein sequences. The ORF8 protein is an accessory protein which is a hypervariable protein and has been associated with SC-2 virulence (Pereira et al., 2020). In the ORF8 dataset, the major number of sequences belonged to the Omicron variant. We identified the most prevalent mutations across the SC-2 variants (Figure 2.34. a). We also hypothesised three PPI binding sites, i.e., siteSK, siteDF1 and siteDF2, on ORF8 (Figure 2.34. b). The sites SK and DF1 possibly interact with host proteins, whereas site DF2 might play a crucial role in the dimerization of ORF8. These sites are suggested as plausible starting points for mutagenesis studies to investigate their biological significance.

(a)



(b)

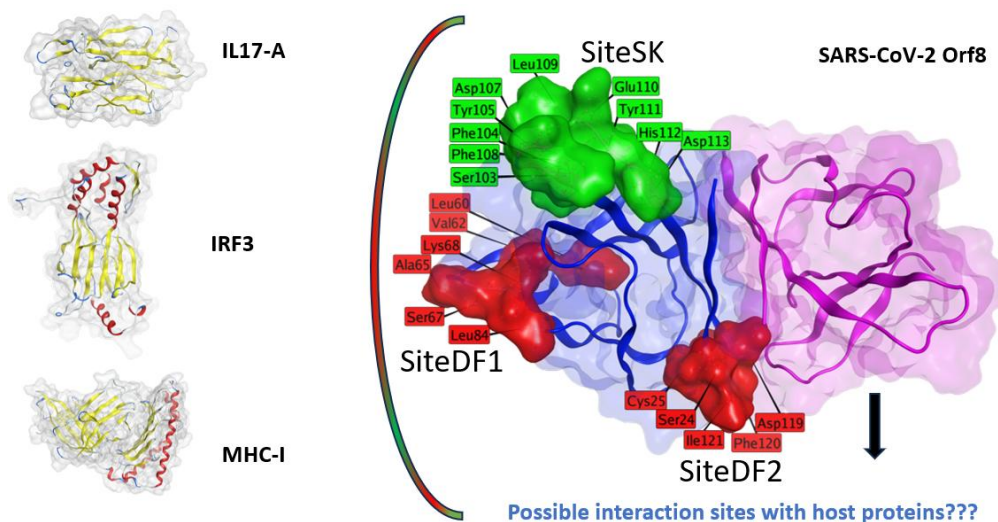


Figure 2.34. (a) SC-2 ORF8 structure with the most prevalent mutations across the SC-2 variants. Alpha (pink), Beta (dark blue), Gamma (cyan), Delta (yellow), Omicron (orange), non-VOCs (purple) and common mutations (red). **(b)** The hypothesised PPI binding sites, i.e., siteSK, siteDF1 and siteDF2, on ORF8 which might interact with host proteins like IL17-A, IRF3, MHC-1 and many more.

The MSA tool benchmarking gave insight into about the time taken by each tool to finish aligning different datasets of sequences. After the benchmarking, we found that Kalign and MAFFT were faster than Clustal Omega. But for the dataset that had 92,989 sequences, Kalign was quicker than MAFFT. For the benchmarking, other MSA tools that work on a different algorithm were also included for a different set of sequences. T-coffee and Muscle MSA tools were still found to be slower than Mafft or Kalign. It was observed that all the outputs produced by the different MSA tools were the same. After the benchmarking, Kalign was chosen for performing the MSA's for the study. When Kalign was used for aligning over a million sequences, it resulted in a segmentation fault. This issue was flagged by me to the developer of Kalign, and he found that the issue was with empty entries that could not be identified by Kalign. On 24th November 2022, he released a new version of Kalign, version 3.3.5, that detects and ignores empty entries in FASTA inputs.

While preparing the dataset, it was found that there were a few sequences that created large gaps within the MSA due to the presence of different amino acids at different positions, which shifted the entire MSA. These sequences impacted the NSP chopping of the MSA with the Python script, thus resulting in an uneven MSA consensus sequence and reference amino acid residue positions in the Excel sheet output file. Hence, it was necessary to remove these sequences from the datasets. As the datasets consist of highly similar sequences (SC-2), the large gaps can occur even with the presence of one dissimilar sequence and can thus impact the chopping of the MSA. The sequence submissions also had a few sequences with unknown AA characters, such as BJOUXZ, which were removed from the datasets, as their presence resulted in warnings while aligning using Kalign.

For the conservational and mutational analysis arbitrary threshold values were selected. Usually, bioinformatics-based studies use an arbitrary selection of thresholds to classify amino acids/nucleotides as being mutated or conserved. For example, (Ou et al., 2022) identified novel mutations in the spike protein of Omicron subvariants if that particular mutation had a frequency of occurrence in > 50 sequences from a total of 52,563 sequences (~0.1). Another study used 2% as the cutoff value for the identification of mutations from a dataset having 383,570 SC-2 genome sequences (Weber et al., 2021). A study by (Perlinska et al., 2022) also used an arbitrary selection of a mutation rate of > 0.002% for identifying frequent mutations from a dataset having 2,610,999 SC-2 PLpro sequences.

Most importantly, the selection of thresholds to consider a range of amino acids as conserved, intermediate variability, and mutated in this study was arbitrary. This issue is prevalent across the broader MSA literature. However, this classification of amino acids is necessary to differentiate between the most prevalent mutants and those with intermediate variability, thereby aiding in the surveillance of mutants. We cannot predict the mutations that the upcoming SC-2 variants will acquire, but exploring these variable amino acid positions can assist us in monitoring their impact on NSPs.

2.11. Conclusions and future directions

- As a significant amount of research has focused on SC-2 spike and protease viral targets, we suggest that the NSPs, which are essential for viral replication, require much more detailed study.
- The bioinformatics pipeline mentioned in this study consisting of sequence alignments, sequence data processing, conservation/mutational analysis and 3D visualisation can be replicated for different projects.
- The conservational analysis on the NSPs highlighted the binding pockets that do not mutate over time and can be targeted for the development of pan SC-2 inhibitors.
- The mutational analysis highlighted the presence of highly variant amino acid positions as well as deletions. These mutations/deletions can be further studied by conducting mutagenesis studies to understand how they improve the overall virus's fitness, help the virus with immune evasion, and affect protein stability.
- Since the sequence dataset size used in the analysis was enormous, this helped in the identification of more populated mutated and conserved amino acid residues.
- In a separate bioinformatics analysis of the SC-2 ORF8 accessory protein, three PPI binding sites, i.e., siteSK, siteDF1, and siteDF2, were hypothesised on ORF8. These sites are suggested as plausible starting points for mutagenesis studies to investigate their biological significance.

Chapter III.

CACHE Challenges

Parts of this chapter have been published in:

Herasymenko O, Silva M, Abu-Saleh AAA, Ahmad A, Alvarado-Huayhuaz J, Arce OEA, Armstrong RJ, Arrowsmith C, Bachta KE, Beck H, Berta D, Bieniek MK, Blay V, Bolotokova A, Bourne PE, Breznik M, Brown PJ, Campbell ADG, Carosati E, Chau I, Cole DJ, Cree B, Dehaen W, Denzinger K, Dos Santos Machado K, Dunn I, Durai P, Edfeldt K, Edwards A, Fayne D, Felfoldi D, Friston K, Ghiabi P, Gibson E, Günther J, Gunnarsson A, Hillisch A, Houston DR, Jensen JH, Harding RJ, Harris KS, Hoffer L, Hogner A, Horton JT, Houliston S, Hultquist JF, Hutchinson A, Irwin JJ, Jukič M, **Kandwal S**, Karlova A, Katis VL, Kich RP, Kireev D, Koes D, Inniss NL, Lessel U, Liu S, Loppnau P, Lu W, Martino S, McGibbon M, Meiler J, Mettu A, Money-Kyrle S, Moretti R, Moroz YS, Muvva C, Newman JA, Obendorf L, Paige B, Pandit A, Park K, Perveen S, Pirie R, Poda G, Protopopov M, Pütter V, Ricci F, Roper NJ, Rosta E, Rzhetskaya M, Sabnis Y, Satchell KJF, Schmitt Kremer F, Scott T, Seitova A, Steinmann C, Talagayev V, Tarkhanova OO, Tatum NJ, Treleaven D, Velasque Werhli A, Walters WP, Wang X, Wells J, Wells G, Westermaier Y, Wolber G, Wortmann L, Zhang J, Zhao Z, Zheng S, Schapira M. CACHE Challenge #2: Targeting the RNA Site of the SARS-CoV-2 Helicase NSP13. *J Chem Inf Model.* **2025**, 14;65(13):6884-6898

Efstathiou C, Zhang Y, **Kandwal S**, Fayne D, Molloy EJ, Stevenson NJ. Respiratory syncytial virus NS1 inhibits anti-viral Interferon- α -induced JAK/STAT signaling, by limiting the nuclear translocation of STAT1. *Front Immunol.* **2024**, 15-1395809

3.1. Introduction

In this chapter, the results obtained from the bioinformatics analysis and the cheminformatics analysis on SC-2 NSP13 and Mac1 domain of NSP3 will be discussed.

3.2. The CACHE challenges

The CACHE (Critical Assessment of Computational Hit-finding Experiments) <https://cache-challenge.org/what-cache> Challenges is an open competition platform to predict small molecules that bind to predefined disease-associated protein targets. The participants develop a computational pipeline to predict small molecules (hits) for the specified target of interest and CACHE will help in providing high quality experimental feedback. The aim of the challenges is to help accelerate early stages of drug discovery. These challenges were initially launched by the Structural Genomics Consortium (SGC) but in 2024 Conscience became the steward for the CACHE challenges. The MDG group participated in CACHE challenge #2 and challenge #3.

- **Challenge #2** (Finding ligands targeting the conserved RNA binding site of SARS-CoV-2 NSP13)
- **Challenge #3** (Finding ligands targeting the macrodomain of SARS-CoV-2 NSP3)

3.3. Conservation analysis used in CACHE challenges

In Chapter II, we discussed the different studies pointing out the conservation between viruses. We also identified conserved binding pockets in the NSPs across the SC-2 variants. In order to develop a pan-coronavirus inhibitor, it is essential to study the sequence as well as structural conservation between the betacoronaviruses. To further expand our conservation analysis study, we also included the group IV viruses in order to identify conserved binding pockets across different viruses. The aim was to identify small molecules, i.e., pan-antivirals, that bind to these conserved binding pockets not only across the betacoronaviruses but also across other group IV viruses. Since we are targeting the most conserved binding pockets, there is the possibility of off-target effects, as there might be human proteins having these similar binding pockets. This concern is later addressed in the chapter. The conservation analysis in this chapter will focus on the SC-2 Mac1 domain of NSP3 and the RNA binding site of NSP13.

3.4. Binding sites used for hit finding

For the **challenge #2**, the RNA binding site of SC-2 **NSP13** was selected by the competition organisers for computational drug design. The predicted compounds were purchased and experimentally tested at CACHE using both enzymatic and binding assays. The SC-2 NSP13 is a viral helicase which belongs to the superfamily 1B and is responsible for the unwinding of double-stranded RNA in a 5' to 3' direction, which is essential for viral replication. The energy from the hydrolysis of nucleotide triphosphate (ATP to ADP) is utilised to catalyse the unwinding of the double-stranded RNA (Tanner et al., 2003). The NSP13 is divided into five domains: an N-terminal zinc-binding domain (ZBD), which coordinates 3 structural zinc ions; a helical “stalk” domain; a beta-barrel 1B domain; and two subdomains, 1A and 2A (RecA), having residues used for the hydrolysis of the ATP to ADP (Newman et al., 2021) (Figure 3.1.).

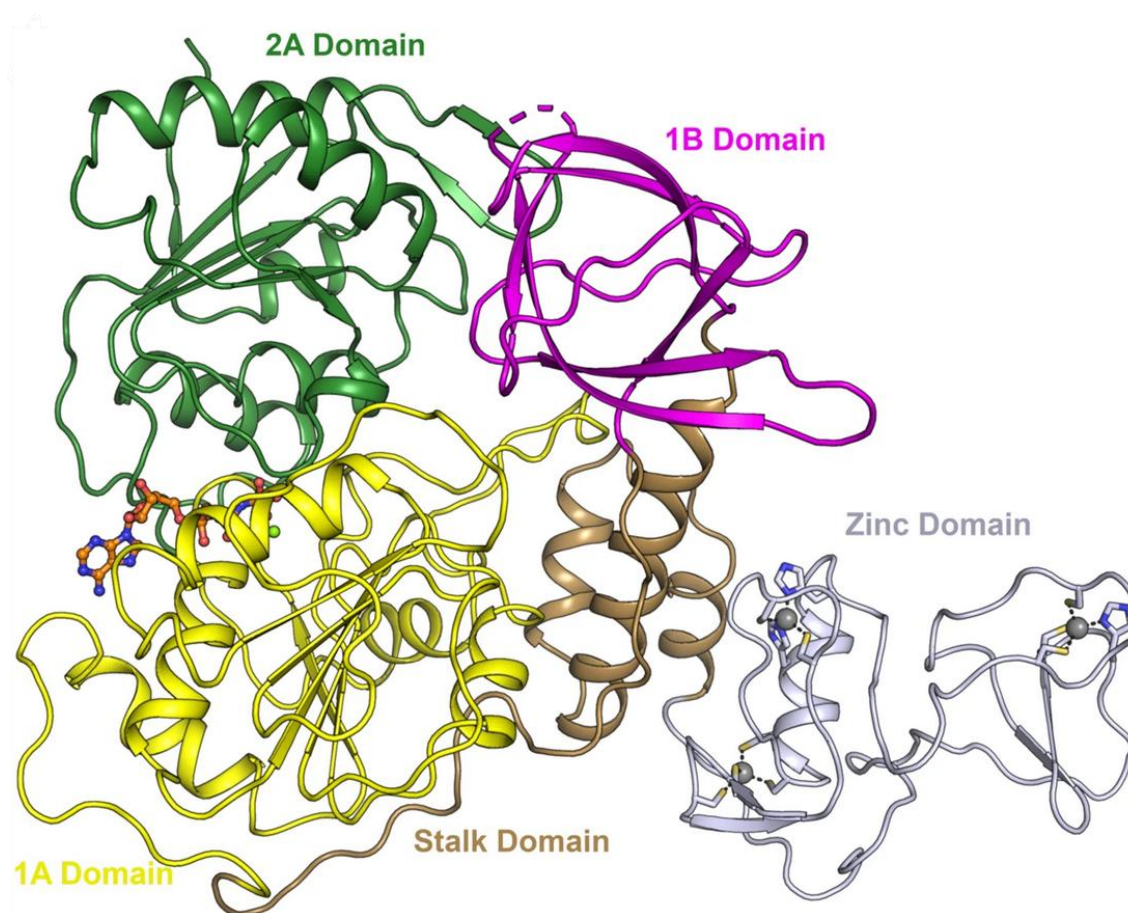


Figure 3.1. Domains of SC-2 NSP13 protein. The 1A domain is represented in yellow, the 2A domain is represented in green, the 1B domain is represented in pink, the zinc domain with zinc ions is represented in grey and ADP is represented in orange. Schematic taken from Newman et al., 2021.

The SC-2 NSP13 protein was selected for the challenge as it was underexplored for the development of small molecule inhibitors during the COVID-19 pandemic. In the past, the helicases have been studied for the development of small molecules targeting either the RNA binding site or the ATP binding site and are a clinically validated target class (Shadrick et al., 2013). For SC-2 NSP13, the RNA binding site is more conserved than the ATP binding site and is druggable, which was reported by the SGC. The RNA binding site of the helicase can be targeted by small molecules but in low micromolar concentrations, e.g., HCV NS3 (LaPlante et al., 2014). The functional activity of SC-2 NSP13 involves RNA binding and helicase-driven RNA unwinding, both of which are essential for viral replication (Sommers et al., 2023). In this challenge, inhibitors will be identified based on experimental evidence that shows compounds binding to the RNA binding tunnel, which means they are interacting with this site. These inhibitors are likely to compete with RNA, thereby blocking its binding and translocation. This disruption might halt helicase function, which in turn impairs viral replication and provides a foundation for antiviral development. The challenge was focused on finding small molecules targeting the RNA-binding site, which had fragments bound to it with PDB codes 5RLH, 5RLZ, 5RML, and 5RMM (Figure 3.2.).

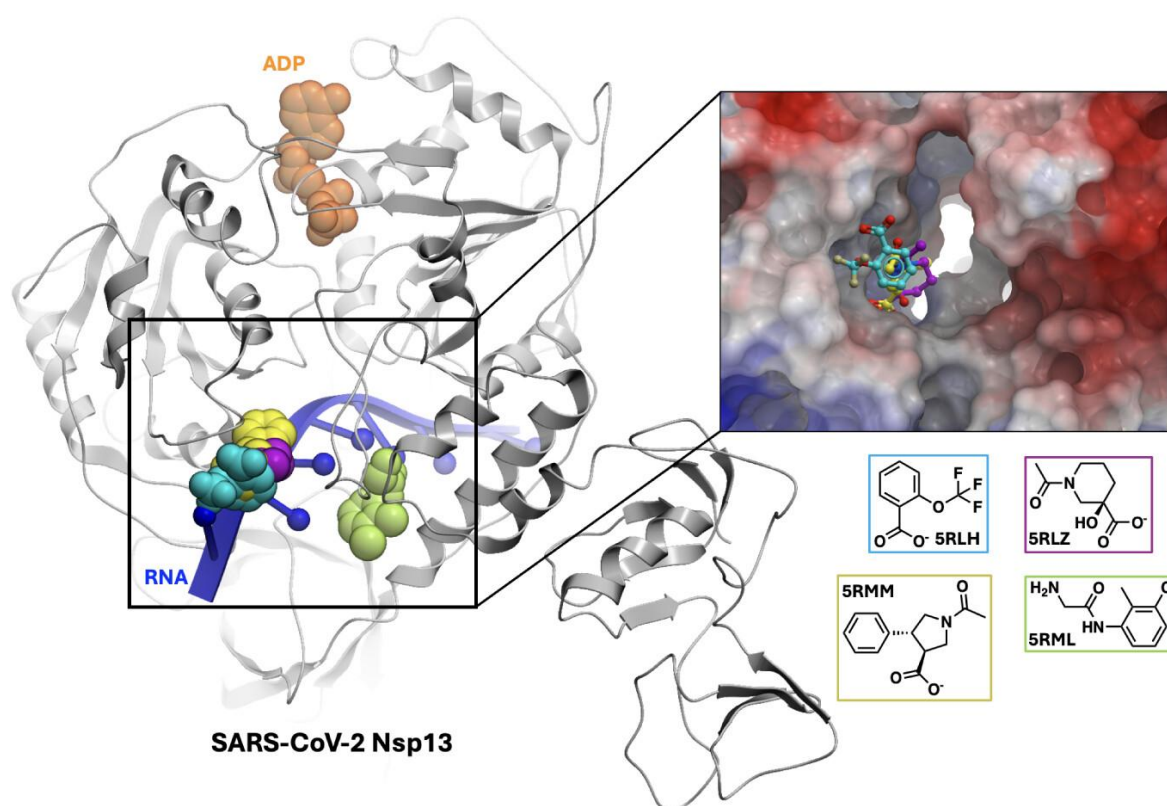


Figure 3.2. SC-2 NSP13 protein (grey) with ATP binding site (hydrolysed to ADP) and the RNA binding site with co-crystallised fragments (5RLH, 5RLZ, 5RML, and 5RMM).

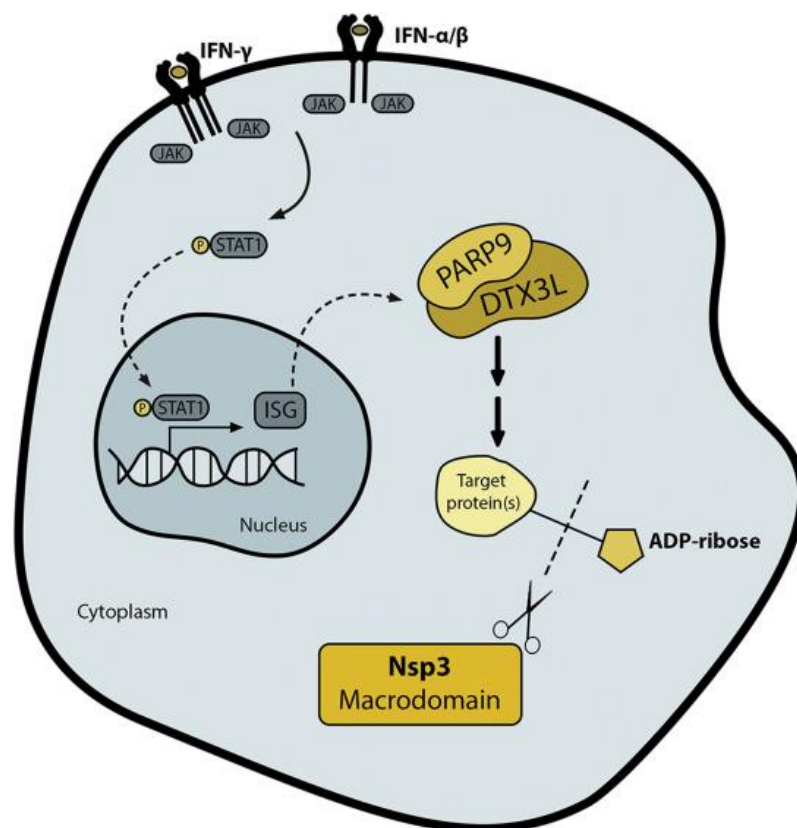
Schematic taken from Herasymenko et al., 2025. Orange: ADP binding site, Dark Blue: RNA binding site, Blue: 5RLH, Purple: 5RLZ, Yellow: 5RMM, Green: 5RML.

For the **challenge #3**, the ADP-ribose binding site of the SC-2 Mac1 domain of **NSP3** was selected by the organisers for computational drug design. The predicted compounds were purchased and experimentally tested at CACHE using binding assays. We covered 50% of the cost for chemical synthesis, and the rest was covered by CACHE, including the biological testing. The interferon response to infections caused by the coronaviruses depends on ADP-ribosylation of the viral proteins. The Mac1 of SC-2 NSP3 counteracts the immune response to viral infection by removing the ADP-ribosylation post-translational modification (Russo et al., 2021) (Figure 3.3. a). The mutational analysis studies have shown that Mac1 is a promising target against SC-1 infections in cellular and animal models (Kuri et al., 2011) (Fehr et al., 2016). The Mac1 domain of SC-2, which is conserved across coronaviruses, is an enzymatic mono-ADP-ribosylhydrolase. The biochemical assays, like gel-based and ELISA-based detection methods for ADP-ribosylated substrates, show that SC-2 Mac1 binds to and removes the mono-ADP-ribose from protein targets through a process known as de-mono-ADP-ribosylation. The structural data like X-ray structures have confirmed the presence of an ADP-ribose binding pocket and its conservation in the Mac1 domains of betacoronaviruses like SC-1 and MERS-CoV. The enzymatic activity of Mac1 counteracts host antiviral ADP-ribosylation and plays an important role in viral pathogenesis, thereby making the SC-2 NSP3 Mac1 domain a validated enzymatic target (Alhammad et al., 2021). The de-ADP-ribosylation of viral proteins can be prevented when inhibitors targeting the SC-2 Mac1 domain compete with ADP-ribose for binding at the active site. Due to the inhibition of SC-2 Mac1 domain by the inhibitors, the host-mediated ADP-ribosylation (addition of ADP-r onto viral proteins) will not be impacted thus enhancing interferon signalling and innate immune defence. A study by (Alhammad et al., 2023) showed that inhibition of SC-2 Mac1 enhances interferon (IFN) and interferon stimulated genes (ISG) responses which leads to rapid viral clearance with minimal pathology in vivo. These effects points to Mac1's role in suppressing host antiviral immunity and support its potential as a therapeutic target.

The ADP-r binding pocket can be subdivided into the adenosine site and the proximal ribose site (Figure 3.3. b). The small molecules were designed to target the binding pocket such that it would compete with its substrate, i.e., the ADP-r. There were high-resolution X-ray

structures of SC-2 Mac1 NSP3 available with its substrate ADP-r, fragments in the binding site (Schuller et al., 2021) and lead-like molecules (Gahbauer et al., 2023).

(a)



(b)

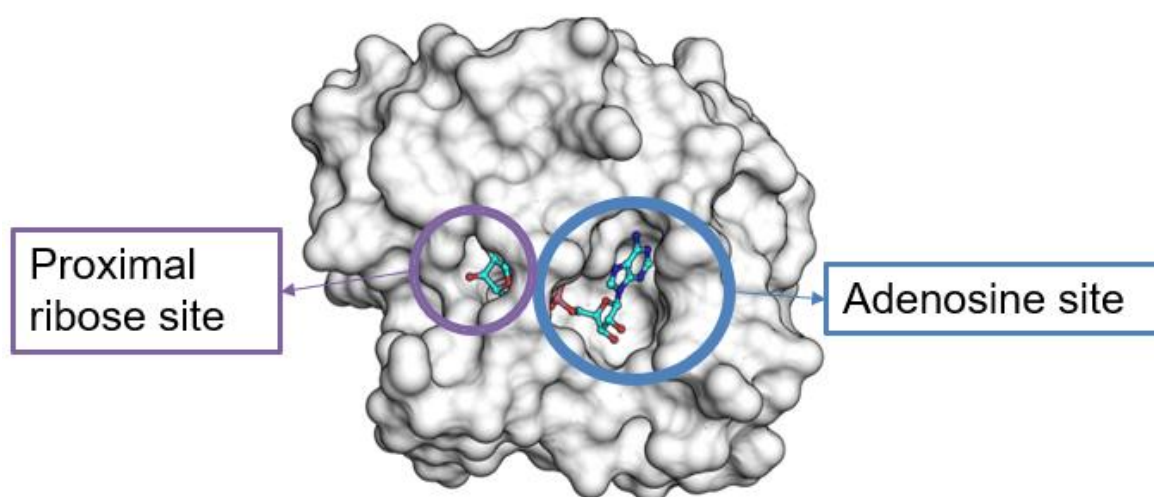


Figure 3.3. (a) Schematic representation of SC-2 NSP3 Mac1 domain removing the ADP-ribosylation post-translational modification that occurs on the viral target protein by PARP9/DTX3L complex, which is activated through interferon (IFN) signalling. The IFN signalling pathway triggers the phosphorylation of STAT1 via JAK kinases, leading to the

expression of interferon-stimulated genes (ISGs) such as PARP9 and DTX3L. The PARP9–DTX3L complex facilitates the ADP-ribosylation of downstream target proteins, a modification that is subsequently removed by the viral NSP3 Mac1 domain. Schematic taken from Russo et al., 2021. (b) ADP-r binding site of SC-2 NSP3 Mac1 domain subdivided into adenosine site (blue) and proximal ribose site (purple).

3.5. CADD techniques

3.5.1. Volume based virtual screening (FastROCS)

Virtual screening (VS) is a crucial step in the CADD workflow, and screening a large chemical library against a binding site is part of the early-stage drug discovery. The OpenEye's FastROCS toolkit, is a novel shape-based VS pipeline. It is based on ROCS – Rapid Overlay of Chemical Structures, which is used to identify similarity between molecules by focusing on their 3D shape as well as chemical features (Rush et al., 2005). The underlying concept is that molecules have similar shape if their volumes overlay and there is dissimilarity between them if their volumes do not overlay. It is a widely used shape-based VS tool which is based on fast Gaussian volume overlay comparison with GPU acceleration. For calculating the shape similarity between two molecules, it utilises Gaussian functions to find and quantify the maximum volume overlap between two molecules (Hawkins et al., 2007) (Grant et al., 1996). In order to improve the efficiency of volume overlap, it only uses equal radii for all heavy atoms but ignores hydrogens. It also utilises only the first-order terms of the shape density function.

ROCS was developed to estimate the similarity in three dimensional shapes of molecules. The similarity between the molecules is defined as:

$$O_{A,B} = \int_V \chi^A(\vec{r}, \vec{q}) \chi^B(\vec{r}, \vec{q}) dV \quad (3)$$

Here, $\chi^A(\vec{r}, \vec{q})$ is the occupation probability (also known as the “characteristic volume”) of molecule A and $\vec{r}$ denotes the position while $\vec{q}$ represents the orientation of the molecule. χ_A is one inside the molecule and zero outside (Rush et al., 2005).

Instead of representing atoms as hard spheres to define their occupation probability within a molecule, Gaussian functions centred at each atom are used to describe this probability more smoothly (Grant & Pickup, 1995). The occupation probability can be represented as:

$$\chi(\vec{r}) = 1 - \prod_{i=1}^{i=N} (1 - g_i(\vec{r})); \quad g_i(\vec{r}) = p_i e^{-\gamma \vec{r}_i^2}; \quad \gamma = \pi \left(\frac{3p}{4\pi R^3} \right)^{2/3} \quad (4)$$

where $\vec{r}_i$ is the distance vector from the atomic center and γ represents the width of the Gaussian function such that the Gaussian volume integral gives the equivalent hard sphere volume of an atom of radius R .

Additionally using $O_{A,B}$, we can estimate the shape distance $D_{A,B}$. The shape distance $D_{A,B}$ is a fundamental quantity and predicts that the shape is an intrinsic property. It is also used to estimate the similarity between molecules.

$$D_{A,B} = \sqrt{O_{A,A} + O_{B,B} - 2O_{A,B}} \quad (5)$$

To calculate the overlap between two molecules, ROCS uses Tanimoto (Rogers & Tanimoto, 1960) and Tversky (Tversky et al., 1977) correlation coefficients as similarity metrics (Figure 3.4.). After the screening, the virtual hits are ranked based on a combo score for Tanimoto similarity (based on volume) and the colour score (based on chemical property).

$$Tanimoto_{AB} = \frac{O_{AB}}{O_A + O_B - O_{AB}} \quad (6)$$

$$Tversky_{AB} = \frac{O_{a,b}}{O_{AB} + \alpha O_A + \beta O_B} \quad (7)$$

Where, O_{AB} represents the overlapping volume between molecules A and B, while O_A and O_B are the total volumes of molecules A and B, respectively. The parameters α and β are weighting factors used in the Tversky index calculation.

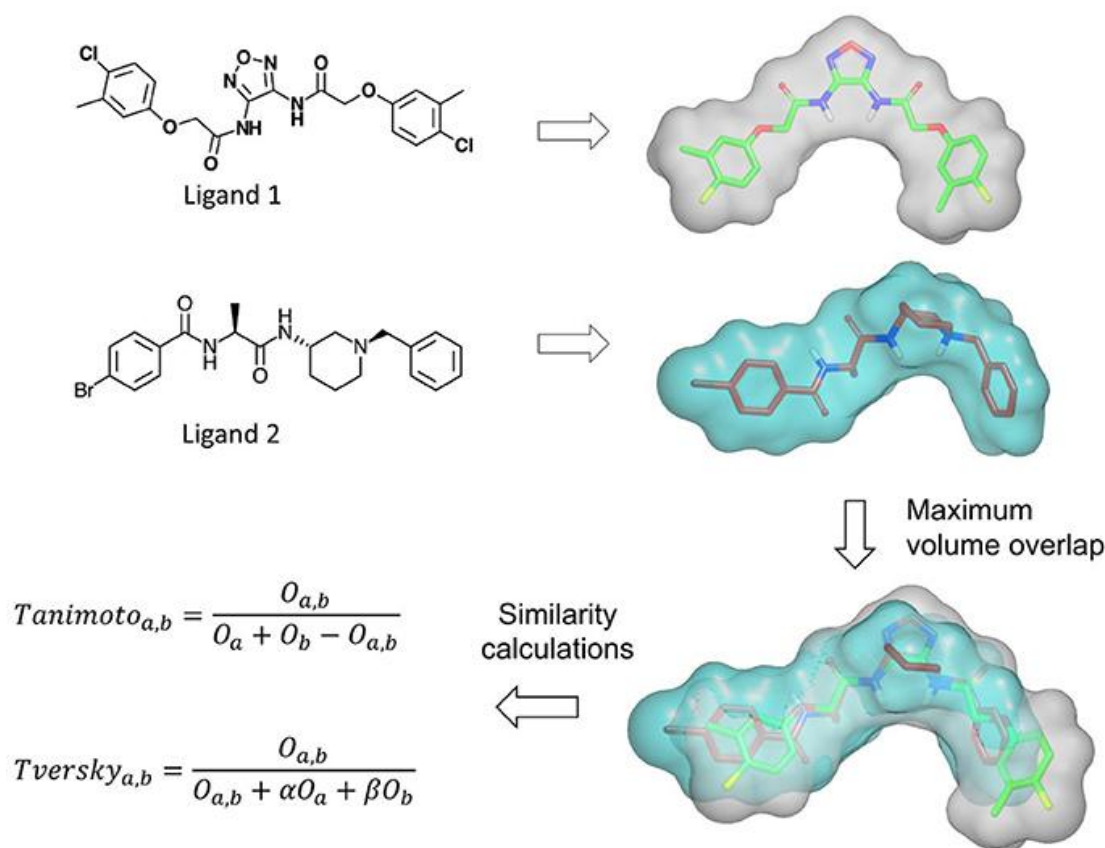


Figure 3.4. ROCS algorithm overview for shape similarity calculation. Example showing structural overlay between Ligand 1 (green) and Ligand 2 (orange) and their volumes are represented in grey and cyan colour respectively. Schematic taken from Kumar et al., 2018.

3.5.2. Pharmacophore based virtual screening

Pharmacophore based VS is based on the theory that having common chemical functionalities, similar 3D spatial arrangements will likely result in similar biological activity on the same target. In a pharmacophoric model, the chemical features of a molecule that can interact with the biological macromolecule are depicted as geometric elements, such as spheres, planes, or vectors. The main pharmacophore feature types are hydrogen bond acceptors (HBAs); hydrogen bond donors (HBDs); hydrophobic areas (H); positively and negatively ionizable groups (PI/NI); and aromatic groups (AR) (Giordano et al., 2022) (Figure 3.5.). The development of the pharmacophore model depends on four stages (Wolber & Sippl, 2015):

1. Selection of active ligands: Selecting a group of compounds (known actives) that bind to the same target and binding site.

2. Conformational analysis: Exploring all possible conformations of each ligand.
3. Assignment of pharmacophoric features: Chemical features responsible for activity.
4. Molecular superimposition: Aligning ligand conformations to create a consensus 3D pharmacophore.

Different pharmacophore modelling tools utilise these steps in order to generate pharmacophore features. The main goal here is to minimize the root-mean-square superposition error between the ligand conformations while maximizing the alignment of pharmacophoric features.

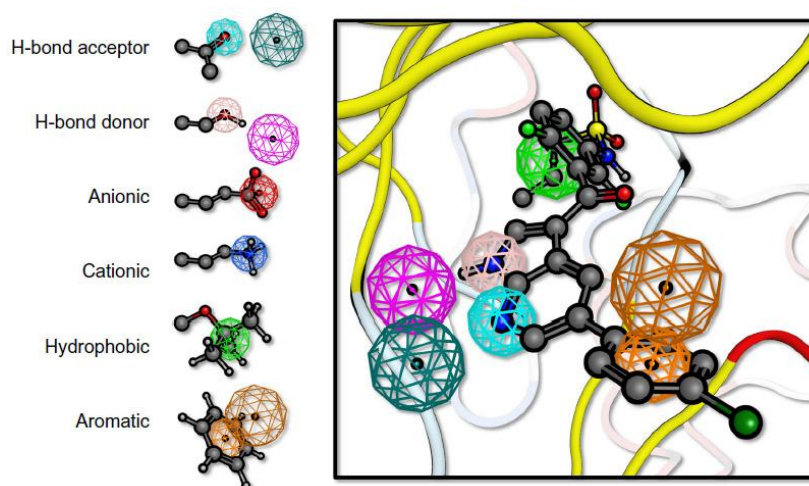


Figure 3.5. Different Pharmacophore features. H-bond acceptors is represented in cyan, H-bond donor is represented in pink, anionic feature is represented in red, cationic feature is represented in blue, hydrophobic feature is represented in green and aromatic feature is represented in orange. Schematic taken from (Qing et al., 2014).

MOE 2022.02 has a consensus-based pharmacophore generation option which depends on the alignment and then superimposition of known active molecules to generate common pharmacophore features. The Mapping of Protein Binding Sites (MoPBS) (Braun & Fayne, 2022) tool, is a pharmacophore-based drug discovery tool. It generates pharmacophore ensembles using K-means clustering of fragment positions in protein binding sites. For generating the pharmacophore features, the protein binding site is flooded with the fragments (Table 3.1.) and after clustering them a pharmacophore feature is assigned.

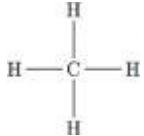
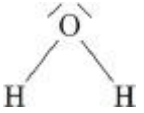
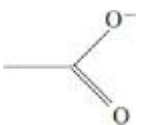
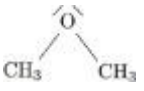
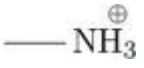
Fragment	Structure	Assigned feature	Assigned point
methane		Hyd	C atom
water		Don, Acc	Don: between H, Acc: O
benzene		Aro	Centroid of ring
acetate ion		Acc	Between O atoms
butane		Hyd	In between C2 and C3
cyclohexane		Hyd	Centroid of ring
dimethylether		Acc, Hyd	Acc: O, Hyd: between CH ₃
ethanol		Don, Hyd	Don: O, Hyd: C2 atom
methylammonium		Don	N atom

Table 3.1. The different fragment types used for binding site flooding with their structure, assigned feature and assigned point. Adapted from (Braun & Fayne, 2022).

3.5.3. FRED molecular docking based virtual screening

Molecular docking tools have been widely used in the drug discovery process for performing VS and lead optimisation (Cross et al., 2009). The docking tools utilise the structure of the protein to perform VS of the molecules present in a certain dataset and examine the interactions between the molecules and the protein binding site. The interactions between the molecule and the binding site are converted into a numerical score with the help of a scoring function which determines the probability that a molecule is a binder. Fast Rigid Exhaustive Docking (FRED) is a molecular docking tool of OpenEye's OEDocking suite. The docking of the molecules in FRED is based on an exhaustive search algorithm which is responsible for searching rotations and translations of each molecule conformer within the binding site. During this search, molecules that have an unrealistic pose or molecules that clash with the protein are filtered out. After the search, the 100 top-ranked molecule poses are subjected to systematic solid-body optimisation. While docking, both protein and conformers are held rigid, but molecule flexibility is taken into consideration as an ensemble of conformers of each molecule is utilised (McGann, 2011) (Figure 3.6.).

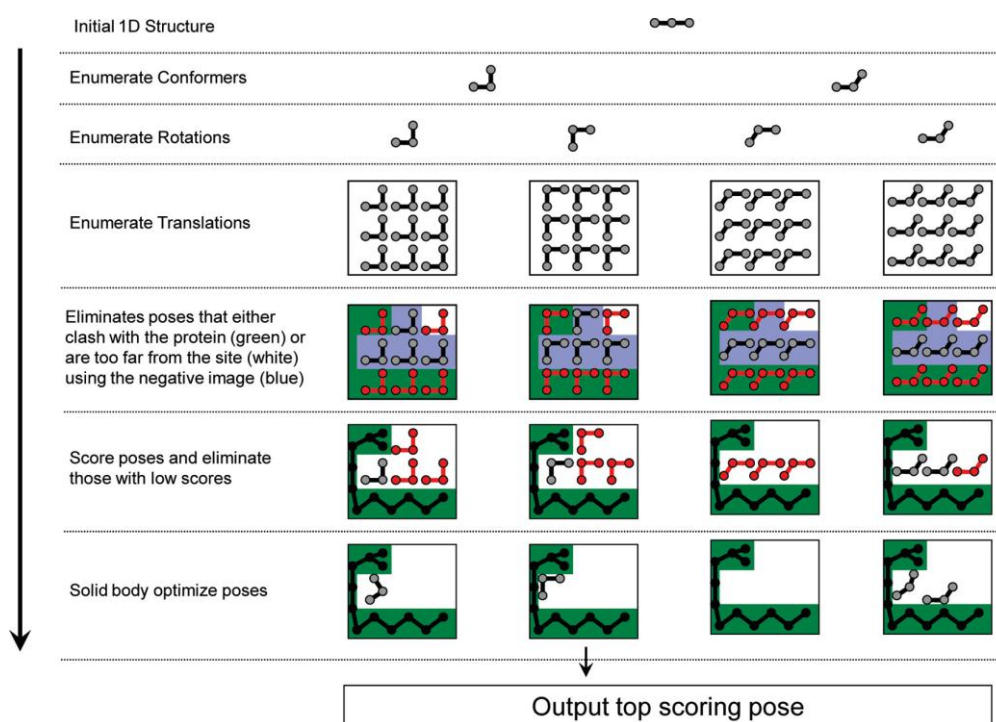


Figure 3.6. FRED docking algorithm workflow. The initial structure is 1D, followed by the generation of conformers, rotations, and translations. The algorithm then eliminated poses that clash with the protein (green) or are too far from the site (white). The scores are generated for the poses, and the ones that have the lowest scores are eliminated. The final poses are optimised and scored. Schematic taken from McGann, 2011.

For the final ranking of the docked poses, the Chemgauss (Eldridge et al., 1997) scoring function is used. The Chemgauss4 scoring function is the most recent version developed by OpenEye and is based on Gaussian-smoothed potentials, which are used to measure how well the ligand poses fit within the binding site. It recognises shape complementarity, hydrogen bonding (protein-ligand or solvent interactions) and metal-chelator interactions. For scoring the poses, the initial potentials are first determined as step functions to describe the interaction of atom pairs as a function of distance (McGann, 2011). These interactions are then mapped onto a 3D grid, and these grid values are convoluted with a spherical Gaussian function.

3.6. Materials and Methods

3.6.1. Bioinformatics materials

- **Databases:** National Center for Biotechnology Information (NCBI- www.ncbi.nlm.nih.gov) and Global Initiative for Sharing All Influenza Data (GISAID- <https://gisaid.org>)
- **Software:** Kalign 3.3.2, Jalview version 2.11.26, Ugene version 41.0, Ubuntu version 20.04.6, BioAider version 1.423, Notepad++ version 8.4.6, Microsoft excel, Windows Subsystem for Linux (WSL), python version 3.8.10, MobaXterm version 21.5, Sublime text editor version 4143
- **Python package:** Seqtk
- **Scripts:** Python script for MSA analysis
- **Operating systems:** Microsoft Windows 11 Education, Linux OS
- **System:** Dell Precision 3650 Tower, Processor- 11th Gen Intel(R) Core™ i9-11900 @ 2.50GHz, 2496 Mhz, 8 Core(s), 16 Logical Processors with NVIDIA graphic card
- **Servers:** Stibbons - AMD Opteron (TM) processor 6276 with 64 cores running on Ubuntu 20.04.6, Irish Centre for High-End Computing (ICHEC supercomputer)

3.6.2. Cheminformatics materials

- **Database:** Protein data bank (PDB - <https://www.rcsb.org>), Enamine database
- **Software:** Molecular Operating Environment (MOE) version 2022.02, OpenEye tools (2022), fastROCS, FRED, MoPBS, DataPype, Make Receptor 4.1.2.2, RDKit, AqSolPred (Sorkun et al., 2020)
- **Scripts:** SVL scripts of MoPBS, DataPype python scripts, substructure search script
- **Operating systems:** Microsoft Windows 11, Linux,
- **System:** Dell Precision 3650 Tower, Processor- 11th Gen Intel(R) Core™ i9-11900 @ 2.50GHz, 2496 Mhz, 8 Core(s), 16 Logical Processors NVIDIA graphic card

3.6.3. Methods for the NSP13 Workflow

(A) Bioinformatics

The bioinformatics workflow of NSP13 is divided into NCBI and GISAID workflow.

NCBI

The MSA and analysis protocol is the same as described in section 2.8. of Chapter II.

- **SC-2 NSP13 comparison with beta-CoVs, group IV viruses and human RNA helicase sequences**

The MSA between SC-2 (YP_009725308) NSP13 and beta-CoVs (SC-1: NP_828870.1 and MERS: YP_009047224.1) was performed using Kalign. The MSA was viewed on Jalview and the MSA was coloured according to the percentage identity. An MSA was also generated of SC-2 NSP13 with group IV viruses helicase protein. The accession ID's were QIM73944.1 (Hepatitis E virus), YP_009507804.1 (Tonate virus), AGE98323.1 (Venezuelan equine encephalitis virus), AAY98760.1 (Rubella virus), NP_690588.1 (CHIKV) and UHH89967.1 (Pirahy virus). A global pairwise sequence alignment was performed between the nsP2 of CHIKV (NP_690588.1:719-957) and SC-2 NSP13 (YP_009725308) using Needle (EMBOSS) on EMBL-EBI (www.ebi.ac.uk). For CHIKV analysis, sequence submissions from NCBI were downloaded with filters: 0 ambiguous characters, complete nucleotides and with name non-structural polyproteins. For analysis msa_result_condence.py python script by Dr. Karsten Hokamp was used. The MSA output files were viewed on Jalview and were coloured on the basis of conservation. Lastly, we wanted to see if there is any presence of helicase binding pocket conservation between SC-2 and human RNA helicases. For this, we

performed MSA of 64 human RNA helicase sequences (Umate et al., 2011) and SC-2 helicase using Kalign to identify AA conservation. The MSA was viewed on Jalview.

GISAID

The SC-2 NSP13 sequence dataset was extracted, analysed and pre-processed as described in **Section 2.8. of Chapter II**. A range of amino acid (AA) sequence lengths (600-606) was selected for performing the MSA.

```
>>> seqtk seq in.fasta -m 600 -M maximum 606 > out.fasta
```

After the extraction of desired AA length, MSA was performed on **10,836,636** sequences with our local server (Stibbons) using Kalign. The MSA was analysed using the output excel file from the python script (msa_condence_result.py). The only difference was that in the NSP reference file only NSP13 sequence was used. BioAider was also used for the mutational analysis.

(B) Cheminformatics

Visualization

The highly conserved residues obtained from the bioinformatics analysis were visualised on SC-2 NSP13 protein using MOE. The conserved residues were selected using the sequence editor and were highlighted (green - highly conserved, white - intermediate conservation, red - mutated). The CHIKV and SC-2 helicase proteins were quick prepped on MOE and structural overlay with match superposition was performed. For overlaying the binding pockets only the residues at the binding site were aligned and superimposed followed by generation of a pairwise RMSD matrix. The same was done for human helicase protein overlay with SC-2 NSP13.

For SC-2 NSP13 X-ray crystal structure analysis (PDB IDs: 5RMM, 5RLH, 5RLZ, 5RML, 6XEZ), the protein overlay was performed with MOE and RMSD was calculated. The X-ray crystal structures 5RMM, 5RLH, 5RLZ and 5RML had fragments co-crystalised within the RNA-binding tunnel. These were used as the starting point of the tiered VS protocol. The fragments 5RMM, 5RLZ and 5RLH were co-crystalized within the same binding pocket and was named as siteA1. Whereas, 5RML had a fragment co-crystalized in a different binding

pocket, this was named as siteA2. Prior to running the overlay, the structures were prepared individually using QuickPrep under Amber10:EHT forcefield on MOE.

Virtual screening workflow

- 1) SiteFinder in MOE was used for identifying putative binding pockets. The dummy atoms were colour coded, and different pockets were generated at the sites where the fragments were co-crystallised. Prior to colour assignment to the pockets, the dummy atoms were converted into carbon and oxygen atoms representing hydrophobic and hydrophilic cavity points respectively. The sdf files were generated for each pocket from the dummy atoms that will be used for fastROCS.
 - Three same site fragment submissions were structurally overlaid, and the structures were quick prepped.
 - Site finder was used to locate dummy atoms on fragment sites and subsets of dummy atoms were created and carbon/oxygen atoms were assigned.
 - The dummy points were selected, and the file was saved in mdb format (file->new->database->save selected only->mdb format). The mdb format was then converted to sdf file.
- 2) The database that we have used for VS is a subset of Enamine, containing ~1.4 billion compounds. The molecules within the database have different heavy atom counts (06-24 HA) and were prepared using open babel with the mmff94smod_NoEstat (this force field variant includes all MMFF94s-Mod terms except Coulomb interaction) forcefield and five conformations of each molecule were generated.
- 3) The first round of VS was the volume query search i.e., by using fastROCS which is a shape-based screening tool. It takes in the sdf of dummy atoms as the query volume and generates a hit list that has molecules mapping the same volume as the dummy atoms.
 - The oepython environment was activated.
 - With fastROCS hits were generated of the sdf files created from each pocket dummy atoms.
 - The database was loaded for the search on Stibbons with ShapeDatabaseServer.py Enamine.oeb

- Search command: ShapeDatabaseClient.py Stibbons:8080 dummy.sdf dummies_hits.sdf
 - The hits were viewed on MOE and an additional column of pocket for which the hits were generated was added.
- 4) The second round of screening was the pharmacophore search. An *in-house* software MoPBS (Braun and Fayne, 2022) was used for generating the pharmacophore features for both A1 and A2 sites. The hits from fastROCS were used as an input for the pharmacophore search.
- MoPBS tool (serial version) was used to create pharmacophore features
 - X-ray fragment 1 (siteA2) - Pharmacophore features were created by flooding the site with 250 fragments. It resulted in 5 features pharmacophore query and the search was carried with all the features. Pharmacophore query with 7 and 6 features were also generated but those features were too small (geometric meshes).
 - X-ray fragments 3 (siteA1) - Pharmacophore query was created by flooding the site with 85 fragments which resulted in 7 features. A partial match of at least 5 features was used as there were 0 hits for 7 and 6 features.
 - To ensure we had the right features, we generated 4 runs for both the sites. The features were overlaid, and most common runs were selected.
 - These features were then used to search the fastROCS hits (to refine the results)
 - In the second approach of screening the features (5 features for both the sites) were used to search only the 24 heavy atoms enamine dataset.
 - The scheme used was PCH_all.
- 5) The next step was then to filter out the molecules from the hit list that had molecular weight > 550 g/mol or Da, SlogP > 5 and pan-assay interference compounds (PAINS-removed hits having 1 value, kept the ones with 0 value). This was a requirement mandated by CACHE.
- MOE database viewer was used to calculate descriptors (NOTE: wait for 10-15 mins for the MOE window to finish calculating)
 - Compute->descriptors->calculate->slogP, weight and PAINS

- 6) Before moving to the final round of VS, the Make Receptor OpenEye tool was used for preparing the receptor for docking. (required for docking)
 - MOE prepared PDB was used
 - Make design Unit
 - Box check- docking volume
 - Choose default values
 - Box shape click on + sign twice
 - Create shape
 - Save it in .oedu

- 7) Omega was used for generating fifty conformations of the molecules present in the hitlist generated after the pharmacophore search.
 - The oepython environment was activated (conda activate oepython) and 50 conformations of hits were generated using omega.
 - `omega2 -in hits.sdf -out 50_confs_hits.sdf -warts false -strict false -verbose true -maxconfs 50 -mpi_np 50 -progress percent -prefix 50_confs_hits_`

- 8) The last round of ultra-refined VS was docking which was performed using FRED. The hit list generated for each site had FRED docking scores and the best/preferred ligand pose.
 - The oepython environment was activated (conda activate oepython).
 - Docking was performed using make receptor and omega conformations as input files.
 - `fred -receptor 3_fragments.oedu -dbase hits_3_fragments_50_confs.sdf -prefix 3frag_test -mpi_np 50 -docked_molecule_file maxhits_3frags__docked.sdf -hitlist_size`
 - `fred -receptor 1fragment.oedu -dbase hits_1_frag_50_confs.sdf -prefix 1frag_test -mpi_np 50 -docked_molecule_file maxhits_1frag__docked.sdf -hitlist_size`

- 9) After the completion of the VS protocol the hits were clustered based on ECFP4 fingerprint using Pipeline Pilot.
 - 1857 compounds were common to 2 of the 3 hitlists
 - 140 clusters from the 2 of 3 compounds and 10 clusters from the 3 of 3 compounds.
 - The compounds that were in more than one hitlist were selected.

- Excel sheet was used for noting the names of the compounds and preparing the sdf files manually. The first molecule from each cluster was selected by compute->sort->select unique entries->1.select cluster->OK. After this go to edit and click on invert selected entries to select the ones that are not the first molecules in each cluster. Go on delete selected. The file thus produced will have the first molecule from each cluster. Look at the excel sheet and match the compounds.
- For selecting the molecules that are present within each cluster and not the first entries go on the find menu select name in the search option->select or->in the search tab paste the ID's of each molecule (ZXXY7 or ZXXY8 or ZXXY9 and so on)->uncheck any of the box from match section->hit select. Do the invert as in the previous step.
- Selected hit lists were combined by import option on MDB viewer->click on + and add files.

Protocol validation

On 16th September 2024, CACHE announced the experimental results of challenge #2. The hit list was acquired from the website and an MS Excel sheet was prepared containing actives and inactives obtained after experimental validation. These molecules were subjected to the tiered VS to identify in which VS step our protocol fails to identify actives.

(C) Experimental validation performed by the CACHE organisation

The **Surface Plasmon Resonance (SPR)** assay was performed using Biacore 8K system (Cytiva) maintained at 20 °C. Biotinylated SC-2 NSP13 was immobilized on flow cell two of a streptavidin-coated sensor chip, with 4900–5100 response units (RU) in accordance with the manufacturer's guidelines. Flow cell one was used as a reference for subtraction considering for each channel. The test compounds were dissolved in 100% DMSO to prepare 10 mM stock solutions, which were then serially diluted (factor 0.5) to yield six concentration points. The compound dilutions were further diluted (1:50) in HBS buffer containing Mg²⁺ (10 mM HEPES, pH 7.4, 150 mM NaCl, 5 mM MgCl₂, 0.03% (v/v) Tween 20), making a final DMSO concentration of 2%. The binding measurements were done using a multicycle kinetic method, with an association phase of 60 s and a dissociation phase of 180 s at a flow rate of 40 µL/min. The equilibrium dissociation constants (K) values were calculated using a steady-state 1:1 binding model implemented in Biacore Insight Evaluation Software (Cytiva) (Herasymenko et al., 2025).

The **ATP hydrolysis** by NSP13 was quantified by identifying the residual ATP levels using a luciferase-based detection assay (Yazdi et al., 2022). The compound inhibition was determined in a 384-well format (final volume of 14 μ L) containing reaction mixtures with 50 mM HEPES (pH 7.5), 5% glycerol, 5 mM magnesium acetate, 5 mM DTT, 0.01% Triton X-100, 0.01% BSA, 0.1 nM NSP13, 3.5 nM 30b PolyT ssDNA, 2.5 μ M ATP, and 2% DMSO. The control wells contained DMSO alone without any compounds. The reactions were started by adding the substrate followed by incubation for 60 minutes at room temperature. Later, 10 μ L of each reaction mixture was transferred to 384-well white plates containing 10 μ L of luciferase reagent (Cat# V6712; Promega, Madison, WI, USA) and incubated for an additional 20 minutes at room temperature. The compounds selected for the dose-response testing were analyzed similarly, and data were processed using GraphPad Prism 9 (Herasymenko et al., 2025).

The solubility of the compounds was investigated using **dynamic light scattering (DLS)**, which detects compound aggregation and laser intensity in solution. The compounds were prepared by serial dilution from DMSO stock solutions. A 50x dilution was also made and added into filtered buffer containing 10 mM HEPES (pH 7.4), 150 mM NaCl, 5 mM $MgCl_2$, and 0.03% Tween20, resulting in a final concentration of 2% DMSO. The prepared samples were transferred into 384-well black, clear-bottom plates (Corning 3540) with 20 μ L per well. The plates were then centrifuged at 3500 rpm for 5 minutes before being analyzed using a DynaPro DLS Plate Reader III (Wyatt Technology) (Herasymenko et al., 2025).

3.6.4. Methods for the NSP3 Workflow

(A) Bioinformatics

The bioinformatics workflow of NSP3 is divided into NCBI and GISAID workflow.

NCBI

The MSA and analysis protocol is the same as mentioned in section 2.8. of Chapter II. The alignment was done on the entire length of NSP3, the Mac1 domain region (167 AA) was extracted from the MSA output file.

- **SC-2 NSP3 comparison with beta-CoVs, group IV viruses and human PARP14 sequence**

MSA between SC-2 (YP_009725299.1) NSP3 and beta-CoVs (SC-1: NP_828862.2 and MERS: YP_009047231.1) was performed using Kalign. The MSA was viewed on Jalview and the MSA was coloured according to the percentage identity. An MSA was also generated of SC-2 NSP3 with group IV viruses Mac1 protein. The accession ID's were QIM73944.1 (Hepatitis E virus), YP_009509052.1 (Tonate virus), AGE98323.1 (Venezuelan equine encephalitis virus), AAY98760.1 (Rubella virus), ATO93732.1 (CHIKV) and UHH89967.1 (Pirahy virus). A global pairwise sequence alignment was performed between the NSP3 of CHIKV (ATO93732.1:1349-1476) and SC-2 NSP3 (YP_009725299.1) using Needle (EMBOSS) on EMBL-EBI (www.ebi.ac.uk). For CHIKV analysis, sequence submissions from NCBI were downloaded with filters: 0 ambiguous characters, complete nucleotides and with name non-structural polyproteins. For analysis msa_result_condence.py python script by Dr. Karsten Hokamp was used. The MSA output files were viewed on Jalview and were coloured on the basis of conservation. Lastly, we wanted to see if there is any presence of ADP-r binding pocket conservation between NSP3 Mac1 domain of SC-2 and human PARP14 protein. For this, we performed a pairwise sequence alignment between human PARP14 sequence (NP_060024.2) and SC-2 NSP3 Mac1 domain sequence using Needle (EMBOSS) on EMBL-EBI (www.ebi.ac.uk) to identify AA conservation.

GISAID

The SC-2 NSP3 sequence dataset was extracted, analysed and pre-processed as mentioned in section **2.8.** of **Chapter II.** A range of AA lengths was selected for performing the MSA.

```
>>> seqtk seq in.fasta -m 1945 -M maximum 1946 > out.fasta
```

After the extraction of desired AA length, MSA was performed on **7,104,871** sequences with our local server (Stibbons) using Kalign. The MSA chopping was performed with a python script (msa_condence_result.py) on our local server Stibbons. The only difference was that in the NSP reference file only NSP3 sequence was used. This alignment was done on the entire length of NSP3, the Mac1 domain region (167 AA) was extracted from the MSA output file.

(B) Cheminformatics

Visualization

The highly conserved residues obtained from the bioinformatics analysis were visualised on SC-2 NSP3 protein using MOE. The conserved residues were selected using the sequence editor and were highlighted (green - highly conserved, white - intermediate conservation, red - mutated). The CHIKV and SC-2 NSP3 Mac1 proteins were quick prepped in MOE and structural overlay with match superposition was performed. For overlaying the binding pockets, only the residues at the binding site were aligned and superimposed followed by generation of a pairwise RMSD matrix. The same was done for human PARP14 protein overlay with SC-2 NSP3 Mac1 protein.

There are X-ray crystal structures available in the PDB with lead-like compounds co-crystallized within the Mac1 domain of NSP3 (Gahbauer et al., 2022) having IC₅₀ values from competitive assays. For SC-2 NSP3 X-ray crystal structure analysis (PDB IDs: 22 PDB structures, most active molecules with IC₅₀ value ≤ 10 μ M, 6WOJ, 7KQP), the protein overlay was performed with MOE and RMSD values were calculated. 6WOJ and 7KQP had the substrate ADP-r co-crystallized within the Mac1 binding pocket. Prior to running the overlay, the structures were prepared individually using QuickPrep with the Amber10:EHT forcefield in MOE.

Virtual screening workflow

- 1) SiteFinder was used in MOE for identifying putative binding pockets. The dummy atoms were colour coded, and different pockets were generated at the sites where the lead-like compounds and ADP-r were co-crystallised. Prior to colour assignment to the pockets, the dummy atoms were converted into carbon and oxygen atoms representing hydrophobic and hydrophilic cavity points respectively. The sdf files were generated for each pocket from the dummy atoms that will be used for fastROCS. In total six pockets were generated.
 - The 22 PDB files were prepared using quick preparation with MOE.
 - The 22 PDB structures were then overlayed using MOE.
 - SiteFinder was then used to locate dummy atoms on overlaid structure.
 - Subsets of dummy atoms were created (pockets 1-4) and carbon/oxygen atoms were assigned for hydrophobic/hydrophilic cavity points.

- Select the dummy points and go to file->new->database->save selected only->mdb format.
 - Save the .mdb files as sdf files.
 - Create sdf file.
 - For pocket5 (closed tunnel protein conformation-6WOJ) and pocket6 (open tunnel protein conformation-7KQP) with ADP, these two PDB submissions were used.
- 2) The database that we have used for VS is a subset of Enamine containing ~1.4 billion compounds. The molecules within the database have different heavy atom counts (06-24 HA) and were prepared using open babel with the mmff94smod_NoEstat (this force field variant includes all MMFF94s-Mod terms except Coulomb interaction) forcefield and five conformations of each molecule were generated.
- 3) The first round of VS was the volume query search i.e., by using fastROCS which is a shape-based screening tool. It takes in the sdf of dummy atoms as the query volume and generates a hit list that has molecules mapping the same volume as the dummy atoms.
- The oepython environment was activated.
 - With fastROCS hits were generated of the sdf files created from each pocket dummy atoms.
 - The database was loaded for the search on Stibbons with ShapeDatabaseServer.py Enamine.oeb
 - Search command: ShapeDatabaseClient.py Stibbons:8080 dummy.sdf dummies_hits.sdf
 - The hits were viewed with MOE and add a column of pocket was added for which the hits were generated.
- 4) The second round of screening was the pharmacophore search. An *in-house* software MoPBS (Braun and Fayne, 2022) was used for generating the pharmacophore features for pockets 2-6. As we had actives for pocket 1, we used consensus Ph4 feature in MOE for generating the pharmacophore features. The hits from fastROCS were used as an input for the pharmacophore search.

Note: Pharmacophore features were generated with actives, dummy atoms and ADP-r.

A. Pocket 1

- Using consensus Ph4 (50% threshold – partial search with Acc at top with large radius 2.9 Å and bottom Acc as essential features) or original. Pharmacophore query: total 9 features with partial search of at least 7.

B. Pocket 2, Pocket 3 and Pocket 4

- Using MoPBS
- Joined all the dummy atoms with single bond.
- Saved the file as a PDB file ready for the preparation step of MoPBS.
- Pocket 2 (Pharmacophore query: total 9 features with partial search of at least 5)
- Pocket 3 (Pharmacophore query: total 9 features)
- Pocket 4 (Pharmacophore query: total 9 features with partial search of at least 5)

C. Pocket 5 with ADP Pocket 6 with ADP

- Using MoPBS with ADP-r as input
 - Pocket 5 (Pharmacophore query: total 7 features)
 - Pocket 6 (Pharmacophore query: total 7 features with partial search of at least 5)
- Used MoPBS software (serial version) to create pharmacophore features for pocket 2 to pocket 4. The features were chosen after multiple runs and the most consistent features were thus selected.
 - The overlaid structure of 22 PDB ID's was used to generate consensus ph4 for pocket1. Click on pharmacophore, click pch_all and then click on consensus. Select the threshold and generate the ph4. The scheme used was PCH_all.
 - These features were then used to search the fastROCS hits (refining the results).
- 5) The next step was then to filter out the molecules from the hit list that had molecular weight > 400 g/mol or Da, SlogP > 3.5, pan-assay interference compounds (PAINS-removed hits having 1 value, kept the ones with 0 value) and carboxylic acid moieties. This was a requirement mandated by CACHE.

- The MOE database viewer was used to calculate descriptors (imp NOTE: wait for 10-15 mins for the MOE window to finish calculating)
 - Compute->descriptors->calculate->slogP, weight and PAINS
 - To remove molecules having carboxylic acid moieties open Database Viewer (DBV) on MOE press Edit | Select to open and attach the Selection application below the DBV header menu. Set Search: Molecule Fields and type the SMARTS match "C(=O)[OH]" for protonated carboxylic acids or "C(=O)[O-]" for deprotonated carboxylic acids. Press the Select button to apply the selection.
- 6) Before moving to the final round of VS the Make Receptor OpenEye tool was used for preparing the receptor for docking. (required for docking)
- MOE prepared PDB
 - Make design Unit
 - Box check- docking volume
 - Choose default values
 - Box shape click on + sign twice
 - Create shape
 - Save it in .oedu
- 7) Omega was used for generating hundred conformations of the molecules present in the hitlist generated after the pharmacophore search.
- Conda activate oepython
 - 100 conformation of hits were generated using omega.
 - `omega2 -in hits.sdf -out 100_confs_hits.sdf -warts false -strict false -verbose true -maxconfs 100 -mpi_np 50 -progress percent -prefix 100_confs_hits_`
- 8) The last round of ultra-refined VS was docking which was performed using FRED. The hit list generated for each site had FRED docking scores and the best/preferred ligand pose.
- Conda activate oepython
 - Docking was performed using make receptor and omega conformations as input files.

- fred -receptor pocket1.oedu -dbase hits_100_confs.sdf -prefix pocket1 -mpi_np 50 -docked_molecule_file maxhits_docked.sdf -hitlist_size
- 9) After the completion of the VS protocol the hits were clustered based on ECFP4 fingerprint using Pipeline Pilot.
- Using Pipeline Pilot on docked hit files
 - 19 compounds were common between 2 hit list belonging to different pockets.
 - Different clusters were prepared for each pocket. Most of the compounds had urea. 10 extra compounds were chosen from pocket 2 as the hits from pocket 4 had globular and bent hits.
 - Selected hit list was combined by import option on MDB viewer->click on + and add files.
- 10) The 22 X-ray crystal structures were used as positive controls for pocket 1 as they were crystalized at that site. In order to check the efficiency of the FRED docking software a positive control test was carried out by creating a haystack consisting of the 22 actives and 1100 inactives (decoys). An *in-house* software called DataPype was used for calculating the benchmarking matrix and checking the efficiency of the docking software. The ROC-AUC values were thus calculated, and values were used to plot a graph on Pipeline Pilot.
- The SMILES of 22 actives were pasted at DUD-E decoy generation website (<https://dude.docking.org/generate>). The site generated 50 decoys of each active.
 - The SMILES were generated after washing the decoys in MOE.
 - Merged the SMILES of 1100 decoys along with 22 actives and generated 100 conformations of the list.
 - Used DataPype and docked the haystack molecules using FRED at pocket 1 site.
- 11) As a validation step for pharmacophore and fastROCS volume query for pocket 1, a positive control test was undertaken. The haystack prepared in step 10 was used as a dataset to screen through the consensus Ph4 pharmacophore query and fastROCS volume query.

Protocol validation

On 10th December 2024, CACHE announced the experimental results of challenge #3. The hit list was acquired from the website and an MS Excel sheet was prepared containing actives and inactives obtained after experimental validation. These molecules were subjected to the tiered VS to identify in which VS step our protocol fails to identify actives.

Hit to lead optimisation

From the CACHE Round 1 results, our two hits were used for substructure searching using RDKit to find structural analogues. The RDKit substructure searching was used to find substructure from the Enamine REAL database. Multiple batch runs were used for the substructure searching by making changes in the core structures of the two hits.

To predict the solubility (LogS) of these analogues AqSolPred ML model was used. We validated its performance by comparing its prediction with MOE predicted LogS values and experimental LogS values. The experimental LogS values were obtained from two datasets, one had 1,128 molecules (<https://github.com/deepchem/deepchem/tree/master>) and the other dataset had 3,942 molecules (<https://github.com/amingh1995/Aqueous-Solubility-Prediction/tree/main>). The co-relation plots were generated using python.

(C) Experimental validation performed by the CACHE organisation

Homogeneous Time-Resolved Fluorescence (HTRF): The binding affinity of test compounds for the SC-2 NSP3 Mac1 domain (residues 205–380) was evaluated using an HTRF-based displacement assay, in which compounds competed with an ADP-ribose–conjugated biotinylated peptide for binding to His₆-tagged protein. The compounds were transferred into ProxiPlate-384 Plus assay plates (PerkinElmer) using an Echo 650 acoustic liquid handler (Beckman Coulter). In a total volume of 20 μ L, containing 12.5 nM NSP3 Mac1 protein, 200 nM peptide ARTK(Bio)QTARK(Aoa-RADP)S (Cambridge Peptides), 0.031 nM terbium-cryptate anti-His monoclonal antibody (HTRF donor, PerkinElmer), and Streptavidin-XL665 (HTRF acceptor, PerkinElmer) in assay buffer (25 mM HEPES, pH 7.0; 20 mM NaCl, 0.05% BSA, 0.05% Tween-20) each reaction was performed. The reagents were introduced using a Multidrop Combi dispenser (ThermoFisher Scientific). The Mac1 protein and peptide mixture were incubated with the test compounds for 30 minutes at room

temperature before adding the HTRF detection reagents, followed by 1-hour incubation at room temperature. The fluorescence signals were recorded on a Synergy H1 microplate reader (Biotek) equipped with an HTRF filter set (A= excitation 330/80 nm, emission at 620/10 nm and B= excitation of 330/80nm, emissions at 665/8 nm). The HTRF ratio values were used to calculate the percentage of ADP-ribose peptide displacement or binding inhibition, with positive controls consisting of 4% DMSO containing both protein and peptide, and negative controls containing 4% DMSO with peptide only (**Herasymenko et al., 2025**).

Surface Plasmon Resonance (SPR): An orthogonal confirmation of compound binding was done using SPR on streptavidin (SA)-coated sensor chips. The experiments were carried out on a Biacore™ 8K instrument (Cytiva) at 20°C. The biotinylated SC-2 NSP3 Mac1 protein (residues 206–374) was immobilized on flow cell two of the SA chip according to the manufacturer's instructions, reaching an immobilization level of approximately 2800–3200 response units (RU). From each channel the flow cell one was left blank and used as a reference for subtraction. The test compounds were prepared as 10 mM DMSO stock solutions and serially diluted (factor 0.5) to generate six concentration points in 100% DMSO. For SPR analysis, each dilution was further diluted 1:25 in HBS-EP+ buffer (10 mM HEPES, pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.05% Tween-20) containing 0.5 mM TCEP, yielding a final DMSO concentration of 4%. The binding measurements were done using a multicycle kinetics protocol, with an association phase of 60 seconds and a dissociation phase of 120 seconds, at a flow rate of 40 µL/min and 20°C. The equilibrium dissociation constants (K_d) were determined by fitting the steady-state response data to a 1:1 binding model using Biacore™ Insight Evaluation software (Cytiva) (**Herasymenko et al., 2025**).

Dynamic Light Scattering: Compound solubility was measured using dynamic light scattering (DLS), which directly detects particle aggregation and laser power in solution. The compounds were prepared at 2.5 mM and 1.25 mM from DMSO stock solutions and subsequently diluted 25-fold into filtered buffer containing 10 mM HEPES (pH 7.4), 150 mM NaCl, 3 mM EDTA, and 0.5 mM TCEP, resulting in a final DMSO concentration of 4%. The prepared samples were transferred into 384-well plates (Corning, Cat# 3540) with 20 µL per well. The plates were centrifuged at 3500 rpm for 5 minutes before the analysis using a DynaPro DLS Plate Reader III (Wyatt Technology) (**Herasymenko et al., 2025**).

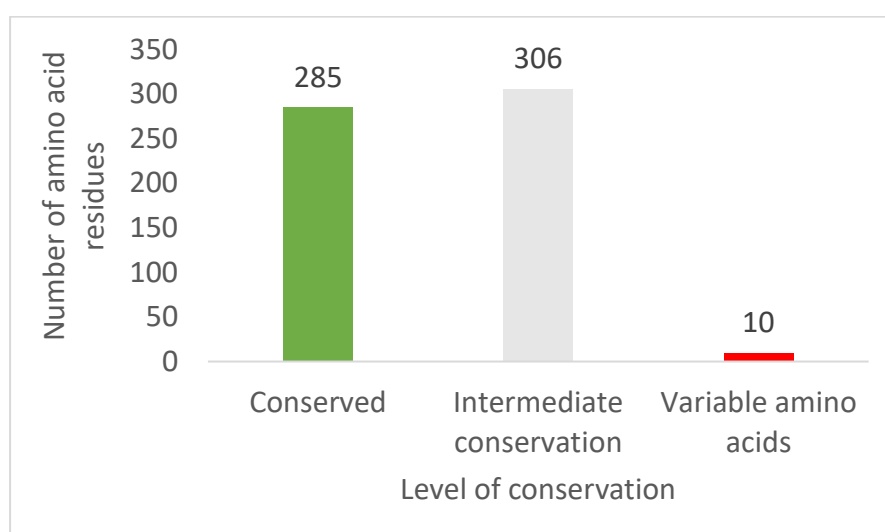
3.7. Results

3.7.1. NSP13 results

(A) NSP13 bioinformatics analysis

From the NCBI dataset analysis on **1,070,751** SC-2 NSP13 sequences (601 AA in length) it was found that there were 285 conserved amino acids, 306 intermediate conservation and top 10 highly variable amino acids (Figure 3.7.) which are discussed in section 2.9.18. of Chapter II. The mutational and conservational threshold selected were 0.437% and 0.002% respectively.

(a)



(b)

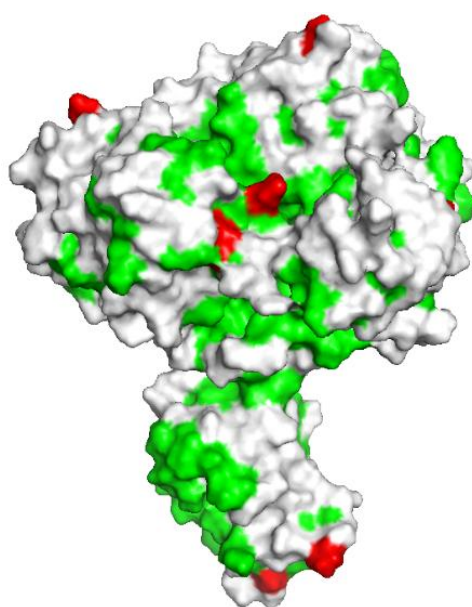


Figure 3.7. (a) Bar graph with number of amino acids belonging to conserved (green), intermediate conservation (white) and variable amino acid (red). The x-axis shows the level of conservation, and the y-axis shows number of amino acid residues. (b) Visualisation of these residues on a 3D protein structure using MOE using same colour scheme as in (a).

As per the mutational analysis on the NCBI dataset using BioAider, it was found that among the top 10 mutations, there were a few mutations that caused changes in amino acid properties. The mutation H164Y (HIS164TYR) led to a change from being polar and positively charged to polar and uncharged. The mutation A296S (ALA296SER) led to a change from being non-polar, uncharged to polar, uncharged. The mutation Q88H (GLN88HIS) led to a change from being polar, non-charged to polar, positive-charged. And lastly, the mutation D260Y (ASP260TYR) led to a change from being polar and negatively charged to polar and uncharged (Figure 3.8.).

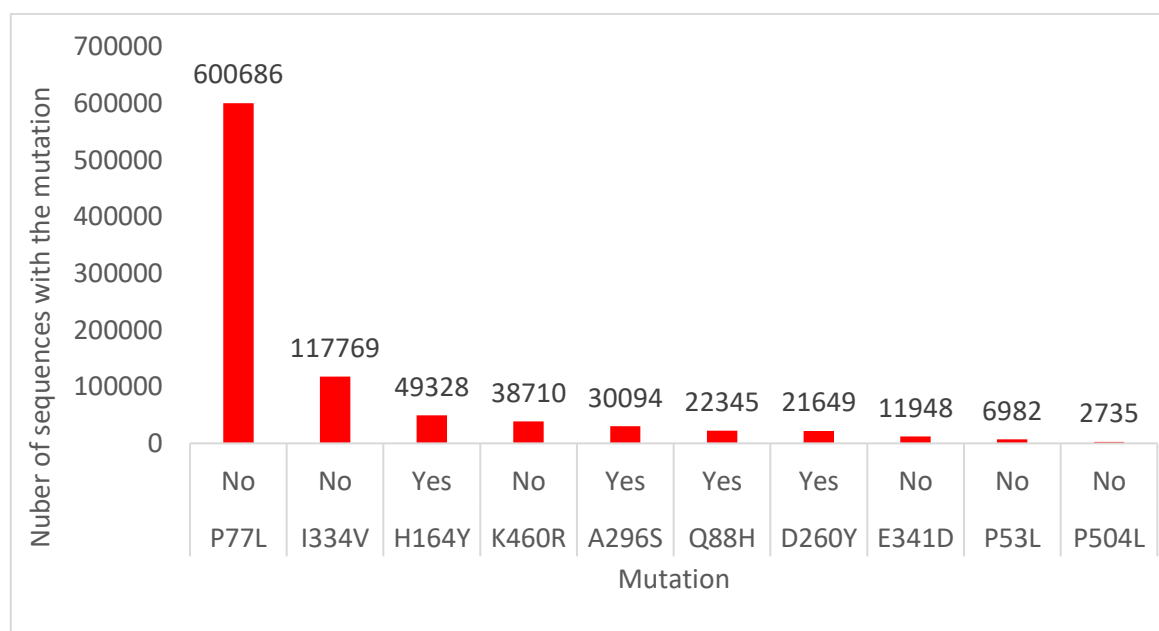


Figure 3.8. Bar graph of the top 10 mutations from the NCBI SC-2 NSP13 sequence dataset. Yes- mutation changed the amino acid properties, No- mutation did not change the amino acid properties. The x-axis shows the type of mutation, and the y-axis shows the number of sequences with the mutation.

In a study (https://apps.thesgc.org/SARSCoV2_pocketome/icm/NSP13.html), the SGC found that the RNA binding site also referred to as the central channel is conserved and has a druggability score of 1.09. The amino acid residues present in the RNA binding

site were found to be highly conserved in the NCBI dataset (Table 3.2.). All of the residues were under the conserved threshold while 378MET was under the intermediate conservation threshold. This RNA binding site/central channel was targeted for the identification of small molecule inhibitors in **challenge #2**.

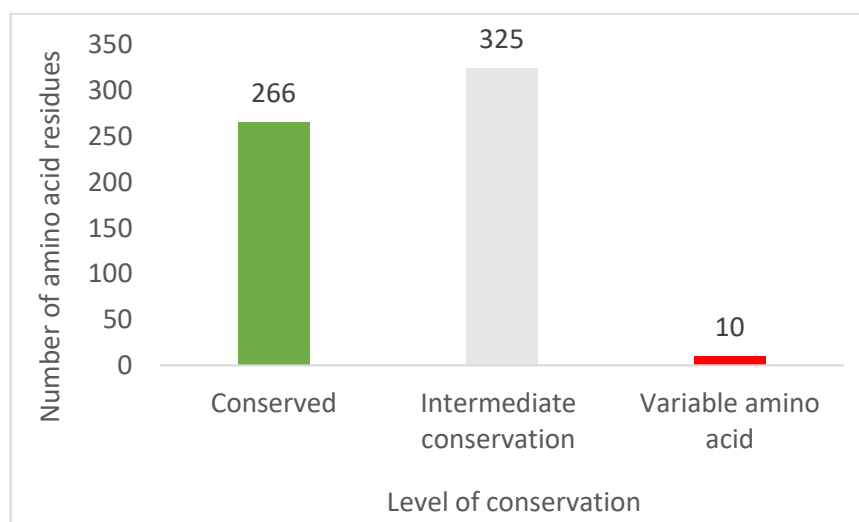
AA position	Amino acid	Number of sequences	Conservation percent	gaps	Mutations	Mutational threshold
177	N	1070744	99.99935	0	S (7)	0.001
178	R	1070750	99.99991	0	Q (1)	0.000
179	N	1070751	100	0		0.000
312	A	1070743	99.99925	0	T (8)	0.001
313	A	1070751	100	0		0.000
316	A	1070749	99.99981	0	T (2)	0.000
374	D	1070751	100	0		0.000
375	E	1070751	100	0		0.000
377	S	1070751	100	0		0.000
378	M	1070717	99.99682	0	I (31); V (3)	0.003
400	G	1070751	100	0		0.000
401	D	1070751	100	0		0.000
404	Q	1070750	99.99991	0	K (1)	0.000
405	L	1070750	99.99991	0	I (1)	0.000
408	P	1070743	99.99925	0	T (8)	0.001
486	S	1070750	99.99991	0	C (1)	0.000
515	Y	1070751	100	0		0.000
516	N	1070751	100	0		0.000
532	T	1070750	99.99991	0	S (1)	0.000
534	D	1070751	100	0		0.000
535	S	1070750	99.99991	0	P (1)	0.000
537	Q	1070750	99.99991	0	P (1)	0.000
538	G	1070746	99.99953	0	C (3); V (2)	0.000
554	H	1070750	99.99991	0	N (1)	0.000
567	R	1070751	100	0		0.000

Table 3.2. List of amino acid residues present at the RNA binding site/central tunnel/central channel of SC-2 NSP13. The conserved residues are highlighted in green, and the intermediate conservation amino acid is highlighted in yellow (NCBI dataset). The AA position column shows the position of amino acid within the NSP13 sequence, amino acid column shows the type of amino acid, number of sequences are the sequences having that exact amino acid, the conservation percent is the percentage of sequences having to amino acid as the reference sequence, gaps column shows if any gaps were inserted, mutations column shows the type of amino acid change along with the number of sequences having that mutated amino acid in brackets and mutational threshold column which shows the percentage of mutation.

From the GISAID dataset analysis on **10,836,636** SC-2 NSP13 sequences (601 AA in length), it was found that there were 266 conserved amino acids, 325 intermediate conservation and top 10 highly variable amino acids (Figure 3.9. a). The mutational and conservational thresholds selected were 0.609% and 0.002%, respectively. The top mutations were PRO77LEU, ARG392CYS, ILE334VAL, LYS460ARG, HIS164TYR, GLU341ASP, ALA296SER, GLU261ASP, SER36PRO and ASP260TYR. Among them, mutations ARG392CYS, GLU341ASP, GLU261ASP and SER36PRO were not identified as top 10 mutations in the NCBI dataset. This result points out the importance of sequence dataset size in the identification of populated mutations. The two datasets (NCBI and GISAID) also share mutations like PRO77LEU, LYS460ARG, HIS164TYR, GLU341ASP, ALA296SER, and ASP260TYR, which were identified in both cases and might hold biological significance.

As per the mutational analysis on the GISAID dataset using BioAider, it was found that among the top 10 mutations, there were a few mutations that caused changes in amino acid properties. The mutation R392C (ARG392CYS) led to a change from being polar and positively charged to polar and uncharged. The mutation H164Y (HIS164TYR) led to a change from being polar, positively charged to polar, uncharged. The mutation A296S (ALA296SER) led to a change from being non-polar, uncharged to polar, uncharged. The mutation S36P (SER36PRO) led to a change from being polar and uncharged to non-polar and uncharged. And lastly, the mutation D260Y (ASP260TYR) led to a change from being polar and negatively charged to polar and uncharged (Figure 3.9. b). The amino acid residues present in the RNA binding site were also found to be highly conserved in the GISAID dataset (Table 3.3.). All of the residues were under the conserved threshold.

(a)



(b)

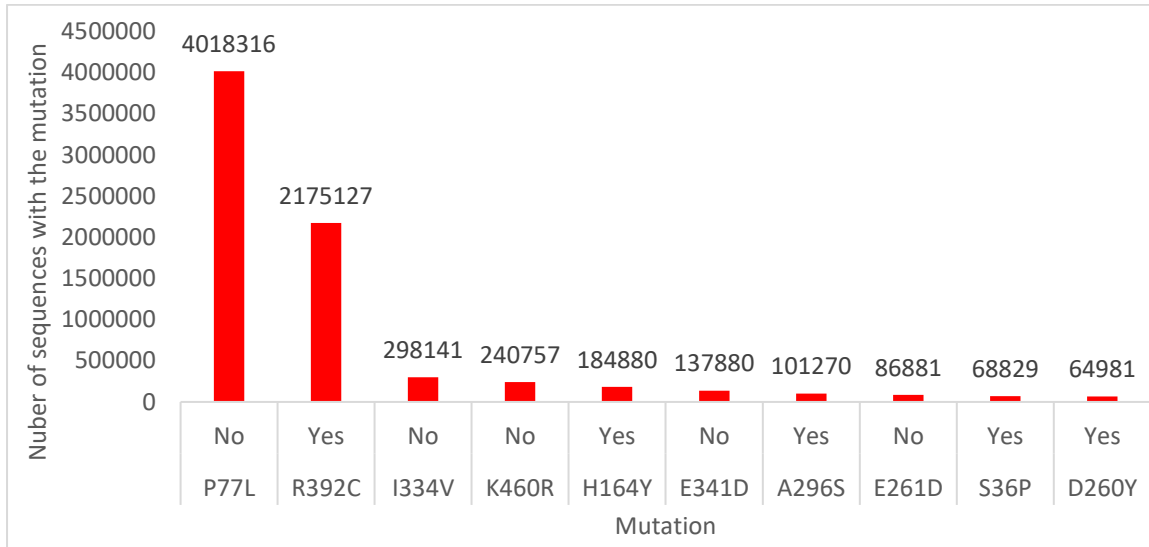


Figure 3.9. (a) Bar graph with number of amino acids belonging to conserved (green), intermediate conservation (white) and variable amino acid (red). The x-axis shows the level of conservation, and the y-axis shows number of amino acid residues (b) Bar graph showing the top 10 mutations from the GISAID SC-2 NSP13 sequence dataset. Yes- mutation changed the amino acid properties, No- mutation did not change the amino acid properties. The x-axis shows the type of mutation, and the y-axis shows the number of sequences with the mutation.

AA position	Amino acid	Number of sequences	Conservation percent	gaps	Mutations	Mutational threshold
177	N	10836500	99.99874	0	S (107); H (3); I,T (1)	0.001
178	R	10836605	99.99971	0	Q (5); L (2)	0.000
179	N	10836603	99.99997	0	K (6); H,D,I (1)	0.000
312	A	10836552	99.99922	0	T (41); S (12); V (4); P (2); D (1)	0.001
313	A	10836568	99.99937	0	D (34); G (5); P (3); T (2)	0.001
316	A	10836524	99.99897	0	T (77); S (8); E,V,P (1)	0.001
374	D	10836605	99.99971	0	G (3); Y (2); E,A (1)	0.000
375	E	10836605	99.99971	0	K (3); G,D (2)	0.000
377	S	10836599	99.99966	0	L (7); P (4); A,T (1)	0.000
378	M	10836395	99.99778	0	I (194); V (20); L (3)	0.002
400	G	10836601	99.99968	0	V (3); E,C (2); N,S,D,A (1)	0.000
401	D	10836605	99.99971	0	N,E,Y (2); A (1)	0.000
404	Q	10836608	99.99974	0	L,K (2)	0.000
405	L	10836604	99.99997	0	I,S (3); V,F (1)	0.000
408	P	10836563	99.99933	0	T (38); A (6); Q (3); L,S (1)	0.001

486	S	10836603	99.9997	0	F (5); Y (3); C (1)	0.000
515	Y	10836605	99.99971	0	C (3); H (2); F,D (1)	0.000
516	N	10836591	99.99958	0	T (17); I (3); K (1)	0.000
532	T	10836606	99.99972	0	I (4); A,S (1)	0.000
534	D	10836605	99.99971	0	Y (3); N,V,A,E (1)	0.000
535	S	10836590	99.99958	0	L (15); A (5); P (2)	0.000
537	Q	10836599	99.99966	0	H (8); L (3); P,K (1)	0.000
538	G	10836598	99.99965	0	C (5); V (4); S,A (2); R (1)	0.000
554	H	10836559	99.99929	0	D (40); Y (9); N (2); R,Q (1)	0.001
566	T	10836603	99.9997	0	N (3); I,S,P (2)	0.000

Table 3.3. List of amino acid residues present at the RNA binding site/central tunnel/central channel of SC-2 NSP13. The conserved residues are highlighted in green (GISAID dataset). The AA position column shows the position of amino acid within the NSP13 sequence, amino acid column shows the type of amino acid, number of sequences are the sequences having that exact amino acid, the conservation percent is the percentage of sequences having to amino acid as the reference sequence, gaps column shows if any gaps were inserted, mutations column shows the type of amino acid change along with the number of sequences having that mutated amino acid in brackets and mutational threshold column which shows the percentage of mutation.

NSP13 conservation across betacoronaviruses

The sequence conservation between the betacoronaviruses (SC-2, SC-1 and MERS-CoV) was strikingly high. The 25 amino acids within the RNA-binding site were 100% conserved (Figure 3.10.) (Table 3.4.).

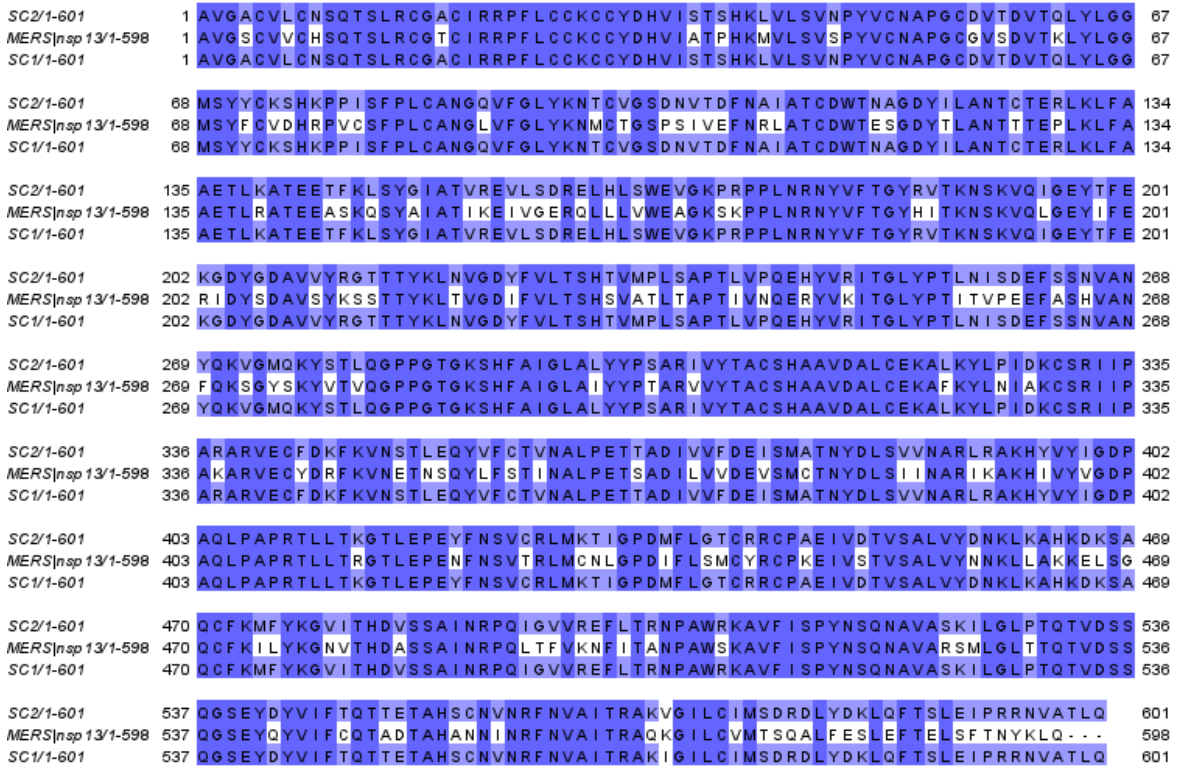


Figure 3.10. MSA of betacoronaviruses (SC-2, SC-1 and MERS-CoV) NSP13 protein. Conserved amino acids are highlighted in purple whereas white colour represents amino acid variability.

position	Amino acid	Number of sequences	Conservation percent
177	N	3	100
178	R	3	100
179	N	3	100
312	A	3	100
313	A	3	100
316	A	3	100
374	D	3	100
375	E	3	100
377	S	3	100
378	M	3	100
400	G	3	100
401	D	3	100
404	Q	3	100
405	L	3	100
408	P	3	100
486	S	3	100
515	Y	3	100
516	N	3	100
532	T	3	100
534	D	3	100
535	S	3	100
537	Q	3	100

538	G	3	100
554	H	3	100
567	R	3	100

Table 3.4. List of amino acid residues present at the RNA binding site/central tunnel/central channel of betacoronaviruses NSP13. The conserved residues are highlighted in green. The AA position column shows the position of amino acid within the NSP13 sequence, amino acid column shows the type of amino acid, number of sequences are the sequences having that exact amino acid, and the conservation percent is the percentage of sequences having to amino acid as the reference sequence (SC-2).

NSP13 conservation across Group IV viruses

An MSA sequence alignment was carried out with group IV helicase sequences. There was not much sequence conservation between them (Figure 3.11.). But a few residues from the RNA binding site were 100% conserved, like ASP374, LEU405, THR532, GLN537, and GLY538 (the residue numbers are as per SC-2 NSP13). Since there was so little sequence conservation observed when performing an MSA of all these group IV viruses against the SC-2 NSP13, we decided to compare these viruses as a part of pairwise sequence alignment (PSA). The PSA was followed by a structural overlay of the group IV viruses and SC-2 NSP13.

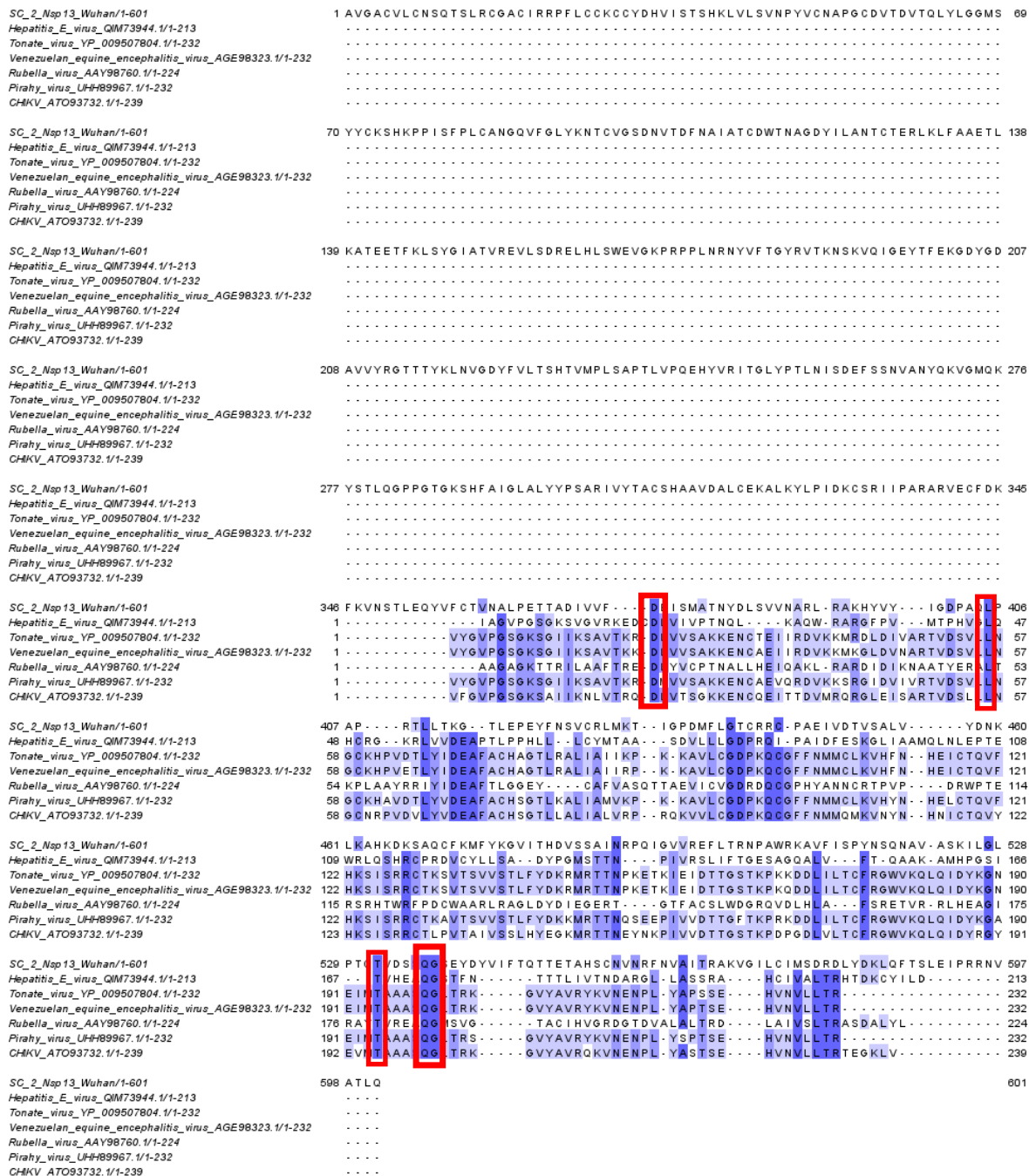


Figure 3.11. MSA of helicase proteins of the group IV viruses (hepatitis E, Tonate virus, Venezuelan equine encephalitis virus, rubella virus, pirahy virus and chikungunya virus) with SC-2 NSP13. The red boxes highlight the conserved amino acids within the RNA binding site. Conserved amino acids are highlighted in purple whereas white colour represents amino acid variability.

Group IV virus sequence and structural conservation to SC-2 NSP13

After comparing sequences and structures of a few group IV helicase proteins against SC-2 NSP13, it was found that Chikungunya virus (CHIKV) nsP2 helicase has sequence/structural

conservation with SC-2 NSP13. Before moving to the comparison, sequence conservation was established for CHIKV nsP2 sequences. A total of 280 CHIKV nsP2 sequences were used for generating the MSA. The mutational and conservational thresholds selected were 2.14% and 0.0%, respectively. Since the dataset was small, large amount of conserved amino acids were found. The top 10 mutations in the dataset were LEU273GLN, LEU16PRO, MET338LYS, SER218THR, HIS374TYR, SER54ASN, HIS130TYR, GLU145ASP, SER405LEU and ARG202THR (Figure 3.12.).

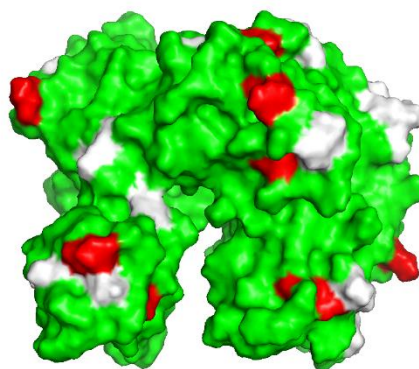


Figure 3.12. Conservational and mutational analysis of CHIKV nsP2 (PDB: 6JIM) (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues.

A PSA analysis was done between the SC-2 helicase (NSP13) and CHIKV helicase (nsP2). The amino acid identity between the two sequences was found to be 8.8% (56 amino acids) and the amino acid similarity was found to be 15.5% (99 amino acids) (Figure 3.13.). The identical amino acids are exactly the same amino acids that are present in the two sequences whereas similar amino acids are the ones that have similar chemical properties and exert similar functions. There were four amino acid positions that were identical and nine amino acid positions that were similar within the RNA binding sites of these proteins (Table 3.5.).

SC-2 AA position	SC-2 amino acid	Type of substitution	CHIKV amino acid
312	A (ALA)	similar AA	I (ILE)
313	A (ALA)	similar AA	T (THR)
374	D (ASP)	identical AA	D (ASP)
375	E (GLU)	identical AA	E (GLU)
377	S (SER)	similar AA	F (PHE)
378	M (MET)	similar AA	A (ALA)
400	G (GLY)	identical AA	G (GLY)
401	D (ASP)	identical AA	D (ASP)
404	Q (GLN)	identical AA	Q (GLN)

405	L (LEU)	similar AA	C (CYS)
486	S (SER)	similar AA	Q (GLN)
515	Y (TYR)	similar AA	S (SER)
516	N (ASN)	similar AA	E (GLU)

Table 3.5. List of similar and identical amino acids present within the RNA binding site of SC-2 and CHIKV helicases. The AA position column shows the position of the amino acid within the NSP13 sequence, the SC-2 amino acid column shows the type of amino acid present in the SC-2 sequence, the type of substitution column shows the similar/identical amino acids between the two sequences, and the CHIKV amino acid column shows the type of amino acid present in the CHIKV sequence.

```
# Length: 637
# Identity: 56/637 ( 8.8%)
# Similarity: 99/637 (15.5%)
# Gaps: 434/637 (68.1%)
# Score: 88.0
#
#=====
```

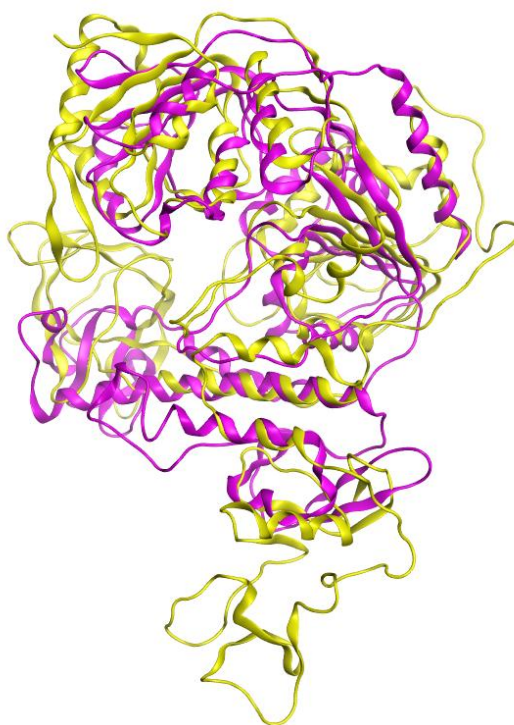
SC-2	1	AVGACVLCNSQTSRLCGACIRRPFLCCKCCYDHVISTSHKLVLSPYVC	50
Chikungunya	1	-----	0
SC-2	51	NAPGCDVTDVTQLYLGGMYYCKSHKPPISFPLCANGQVFGLYKNTCVGS	100
Chikungunya	1	-----	0
SC-2	101	DNVTDNFNAIATCDWITNAGDYILANTCTERLKLFAAETLKATEETFKLSYG	150
Chikungunya	1	-----	0
SC-2	151	IATVREVLSDRELHLSWEVGKPRPPLNRNVFTGYRVTKNSKVQIGEYTF	200
Chikungunya	1	-----	0
SC-2	201	EKG DYGD AVVYRGTTTYKLNVGDYFVLTSHTVMPLSAPTLVPQEHYVRIT	250
Chikungunya	1	-----	0
SC-2	251	GLYPTLNISDEFSSNANYQKVGKYSTLQGGPGTGKSHFAIGLALYYP	300
Chikungunya	1	-----VFVGPVSGSKSAIKNLV----	17
SC-2	301	SARIVYTA-----CSHAADVLCALKYLPIDKCSRIIPARARVECFDK	345
Chikungunya	18	TRQDLVTSKGKENCQEITTDVNRQGLE-----ISART-----VDS	53
SC-2	346	FKVNSTLEQYVFCTVNALPETTADIVFDEISMATNYDLSVNNARLRAKH	395
Chikungunya	54	LLNG-----C-----NRPVDVLYVDEAFACHSGTLALIALVRPRQ	90
SC-2	396	VYVI-GDPAQLPAPRTLLTKGLEPEYFNSVCRIM--KTIGPDMFLGTCR	442
Chikungunya	91	KVVLCDPKQC---GFFNMVQMKVNNHNICTQVYHKSIS-----R	128
SC-2	443	RCPAEIVDTVSALVYDNKLAHK-----DKSAQCFK	473
Chikungunya	129	RCTLPVTAIVSSSLHYEGKMRITNEYNKPIVDTTGSTKPDGDLVTCFR	178
SC-2	474	-----MFYKG--VITHDVSSAINRPQIGVVREFLTRNPAWRKAVFISP	514
Chikungunya	179	GWVKQLQIDYRGHEVMTAAASQGLTRKGVYAVRQKVNENP-----LYAST	223
SC-2	515	YNSQNAVASKIILGLPTQTVDSSQGEYDYVIFTQTETAHSCNVRFNVA	564
Chikungunya	224	SEHVNVLTRTEGKLV-----	239
SC-2	565	ITRAKVGILCIMSRLDYDKLQFTSLEIPRRNVATLQ	601
Chikungunya	240	-----	239

Figure 3.13. PSA between SC-2 NSP13 and CHIKV nsP2. (|) indicate identical amino acids, (:) indicate high similarity and (.) indicate weak similarity.

The structural overlay of SC-2 NSP13 and CHIKV nsP2 showed a good overall structural overlay with an RMSD of 2.6 Å (Figure 3.14. a). The structural and sequence overlay was

then performed on the two binding sites of the helicase protein. The ADP-binding pocket of the two helicases was found to be highly conserved, with an overall RMSD of 9.8 Å (Figure 3.14. b). The overlay showed only one amino acid position deviating with an RMSD of 41.2 Å. The RNA-binding site between the two proteins showed a good sequence and structural overlay with an overall RMSD of 6.54 Å (Figure 3.14. c). We also investigated the structural conservation between the two proteins, considering the binding site of the fragments used in the VS workflow. The sequence and structural overlay of the site where only one fragment was co-crystallised, also referred to as site A2, had a good overlay with an RMSD of 4.1 Å (Figure 3.14.d). In contrast, the site where three fragments were co-crystallized, known as site A1, exhibited a poor overlay with an RMSD of 17.6 Å (Figure 3.14. e). The RMSDs are high in these cases because both sequence and structural aspects were taken into account during the overlays.

(a)



(b)

Pairwise RMSD Matrix:

Chains	1	3
1:7RDY.E		9.89
3:6JIM.A	9.89	

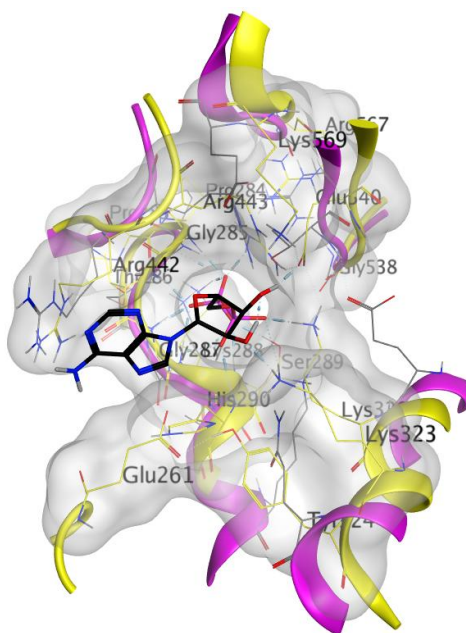
Overall RMSD: 9.887 Å

1: 7RDY.E REPLICATION/TRANSCRIPTION
vs.

3: 6JIM.A VIRAL PROTEIN/RNA

RMSD = 9.887 Å

ALA 174	-	GLU 261	:	41.229	!
VAL 187	-	PRO 283	:	1.542	
PRO 188	-	PRO 284	:	1.570	
GLY 189	-	GLY 285	:	1.852	
SER 190	-	THR 286	:	2.205	!
GLY 191	-	GLY 287	:	1.817	
LYS 192	-	LYS 288	:	1.925	
SER 193	-	SER 289	:	1.897	
ALA 194	-	HIS 290	:	1.732	
ASP 220	-	LYS 320	:	1.292	
ARG 223	-	LYS 323	:	1.444	
GLN 224	-	TYR 324	:	1.558	
ARG 311	-	ARG 442	:	2.728	!
ARG 312	-	ARG 443	:	2.243	!
GLY 384	-	GLY 538	:	0.693	
THR 386	-	GLU 540	:	2.099	!
ARG 416	-	ARG 567	:	1.409	
GLU 418	-	LYS 569	:	2.748	!



(c)

Pairwise RMSD Matrix:

Chains	1	3
1:7RDY.E		6.54
3:6JIM.A	6.54	

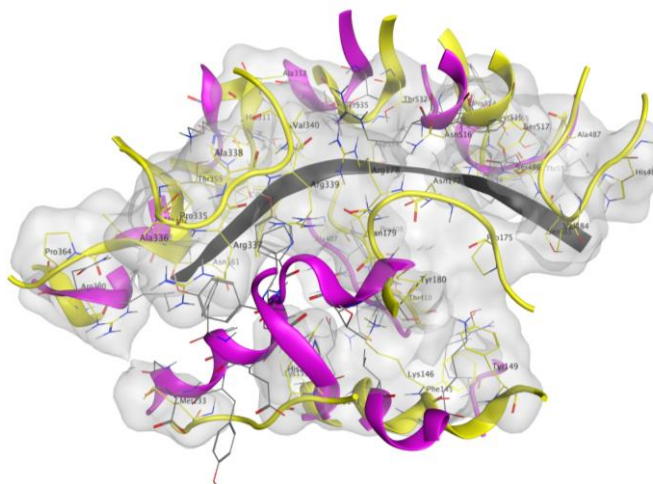
Overall RMSD: 6.541 Å

1: 7RDY.E REPLICATION/TRANSCRIPTION
vs.

3: 6JIM.A VIRAL PROTEIN/RNA

RMSD = 6.541 Å

VAL 92	-	LYS 139	:	2.499	!
ARG 96	-	GLU 143	:	2.958	!
PHE 98	-	PHE 145	:	3.025	!
VAL 99	-	LYS 146	:	3.132	!
LYS 102	-	TYR 149	:	4.370	!
ALA 112	-	PRO 175	:	10.781	!
ASN 114	-	ASN 177	:	14.552	!
THR 115	-	ARG 178	:	14.612	!
ASP 116	-	ASN 179	:	11.197	!
GLU 117	-	TYR 180	:	8.771	!
ASP 155	-	HIS 230	:	12.354	!
ASN 158	-	MET 233	:	17.993	!
GLY 209	-	CYS 309	:	4.094	!
LYS 210	-	SER 310	:	4.581	!
LYS 211	-	HIS 311	:	3.589	!
GLU 212	-	ALA 312	:	3.462	!
THR 233	-	THR 359	:	3.811	!
ASP 235	-	ASN 361	:	3.359	!
SER 236	-	ALA 362	:	3.958	!
LEU 237	-	LEU 363	:	4.135	!
LEU 238	-	PRO 364	:	7.540	!
ALA 256	-	MET 378	:	2.342	!
LEU 268	-	ARG 390	:	5.504	!
PHE 286	-	ALA 407	:	1.833	
PHE 287	-	PRO 408	:	2.510	!
ASN 288	-	ARG 409	:	1.988	
MET 289	-	THR 410	:	2.120	!
CYS 359	-	PRO 514	:	3.223	!
PHE 360	-	TYR 515	:	2.687	!
ARG 361	-	ASN 516	:	3.044	!
GLY 362	-	SER 517	:	3.470	!
THR 378	-	THR 532	:	2.618	!
ALA 380	-	ASP 534	:	2.545	!
ALA 381	-	SER 535	:	2.728	!
ASN 400	-	THR 552	:	4.946	!
PRO 401	-	ALA 553	:	4.335	!
LEU 402	-	HIS 554	:	3.905	!
TYR 403	-	SER 555	:	1.628	



(d)

Pairwise RMSD Matrix:

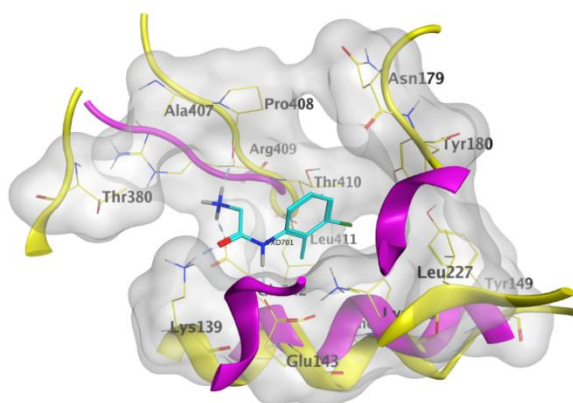
Chains	1	3
1:5RML.A		4.17
3:6JIM.A	4.17	

Overall RMSD: 4.173 Å

1: 5RML.A VIRAL PROTEIN
vs.
3: 6JIM.A VIRAL PROTEIN/RNA

RMSD = 4.173 Å

VAL 92	-	LYS 139	:	1.305	
GLU 95	-	GLU 142	:	1.118	
ARG 96	-	GLU 143	:	0.997	
PHE 98	-	PHE 145	:	2.374	!
VAL 99	-	LYS 146	:	2.433	!
ASP 116	-	ASN 179	:	7.840	!
GLU 117	-	TYR 180	:	6.742	!
LEU 152	-	LEU 227	:	6.509	!
HIS 258	-	THR 380	:	5.726	!
PHE 286	-	ALA 407	:	4.572	!
PHE 287	-	PRO 408	:	3.948	!
ASN 288	-	ARG 409	:	1.658	
MET 289	-	THR 410	:	2.015	!
MET 290	-	LEU 411	:	1.714	



(e)

Pairwise RMSD Matrix:

Chains	1	5
1:5RLZ.B		17.68
5:6JIM.A	17.68	

Overall RMSD: 17.678 Å

1: 5RLZ.B VIRAL PROTEIN
vs.
5: 6JIM.A VIRAL PROTEIN/RNA

RMSD = 17.678 Å

ALA 112	-	PRO 175	:	12.326	!
ASN 114	-	ASN 177	:	16.305	!
ASP 116	-	ASN 179	:	13.236	!
GLU 117	-	TYR 180	:	12.062	!
THR 345	-	SER 485	:	25.199	!
GLY 346	-	SER 486	:	20.848	!
SER 347	-	ALA 487	:	21.315	!
TYR 371	-	PRO 514	:	16.845	!
ARG 372	-	TYR 515	:	18.487	!
GLY 373	-	ASN 516	:	18.166	!
TYR 374	-	SER 517	:	17.705	!
GLY 389	-	THR 532	:	14.787	!

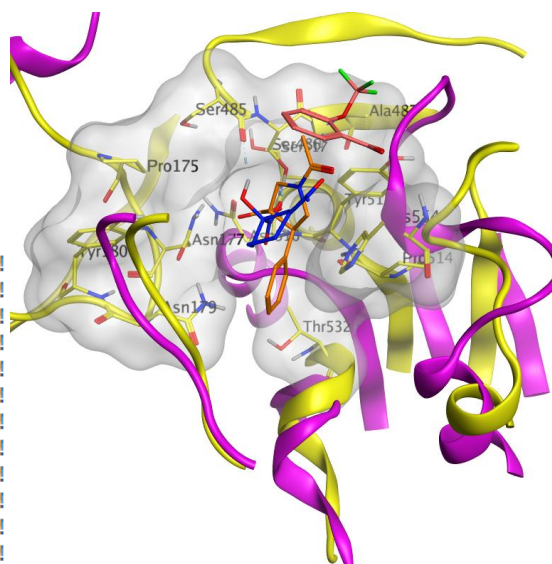
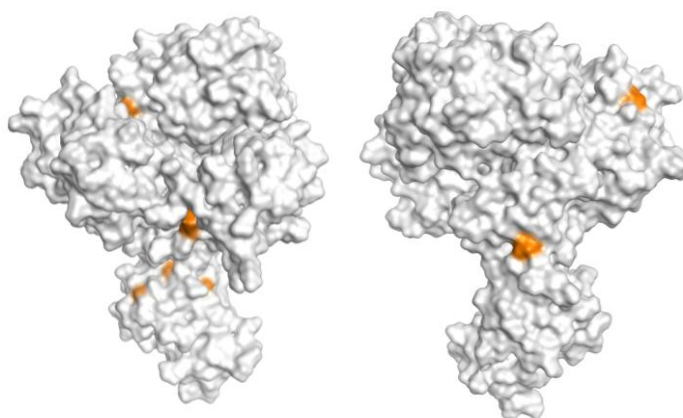


Figure 3.14. (a) Structural overlay between SC-2 NSP13 (yellow, PDB: 5RML) and CHIKV helicase (pink, PDB: 6JIM). (b) Structural overlay of the ADP-binding (black) site between SC-2 NSP13 (yellow, PDB: 7RDY) and CHIKV helicase (pink, PDB: 6JIM) along with pairwise RMSD matrix. (c) Structural overlay of the RNA-binding (black) site between SC-2 NSP13

(b)



(c)

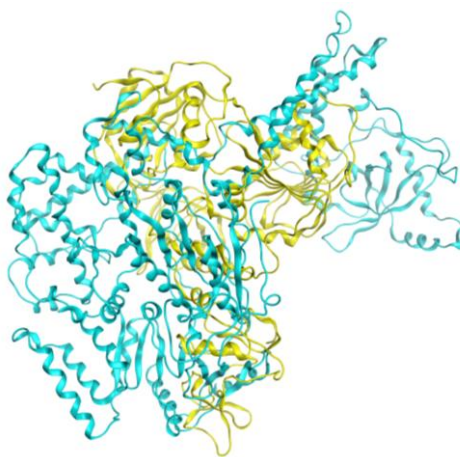


Figure 3.15. (a) A subsection (1-111 residues of SC-3 NSP13) of MSA between NSP13 of SC-2 and 64 human RNA helicases. Colour scheme: percentage identity (indigo colour to white as identity decreases). (b) Residues common (orange colour) between NSP13 of SC-2 and 64 human RNA helicases on NSP13 protein structure (PDB ID: 5RML) with two different orientations (front and back). (c) Structural overlay between human RNA helicase (cyan) (PDB: 7QE0) and SC-2 NSP13 (PDB: 5RLZ) (yellow).

(B) NSP13 cheminformatics analysis

Protein overlay and binding site analysis

The CACHE #Challenge2 was focussed on targeting the RNA-binding tunnel of SC-2 NSP13. There were four PDB submissions available with fragments co-crystallised at different sites that were aligned and overlaid using MOE to check similarity. The average RMSD value of the overlay was 1.20 Å, suggesting that there was very little difference between these X-

ray structures (Figure 3.16.). Residue positions 204-206, that are not within the targeted sites, deviated with an RMSD value reaching 4 Å.

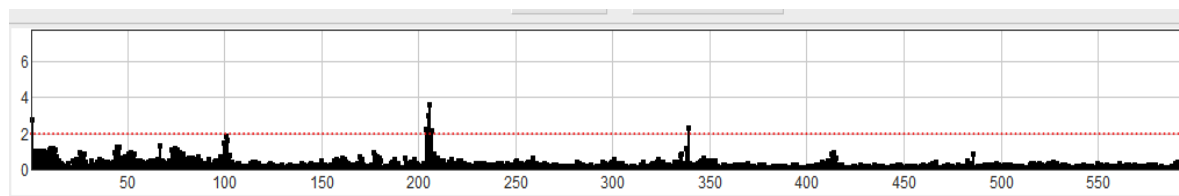


Figure 3.16. RMSD plots (y-axis) of each amino acid residue position (x-axis). The average RMSD is plotted in red dotted line.

The site where three fragments were co-crystalised was named as siteA1 whereas, the site that had only one fragment was named as siteA2 (Figure 3.17). The binding site was observed using the SiteView option on MOE. The AAs present within range of 4.5 Å of the binding pocket were noted. MOE SiteFinder helps in predicting possible binding pockets by assigning dummy atoms that represent hydrophilic and hydrophobic cavity points. SiteFinder was able to predict the sites where the fragments were co-crystalised (Table 3.6.).

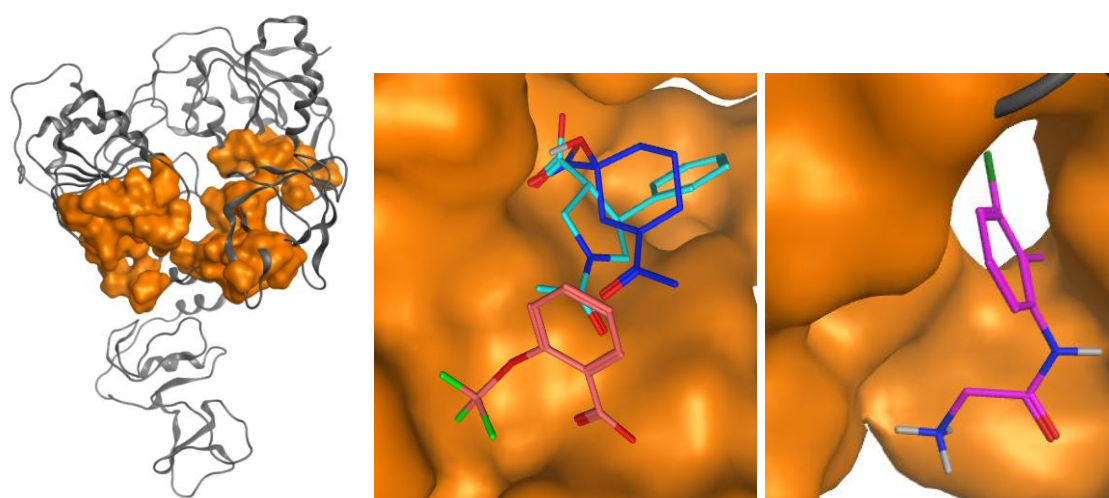


Figure 3.17. The RNA-binding tunnel (orange colour) of SC-2 NSP13 and the different sites that are targeted in this study. SiteA1 has three fragments (peach, blue and cyan) and siteA2 has the single fragment (pink).

Site	PDB ID's	Motif Name	Binding pocket 4.5 Å	Site Finder
RNA Binding Sites	5RMM, 5RLH, 5RLZ	SiteA1	Asn177, Asn179, His482, Ser485, Ser486, Ala487, Pro514, Tyr515, Asn516, Thr532 (motif V), Asp534 (motif V), Thr550, Thr552, Ala553, His554, water	LEU176 ASN177 ARG178 ASN179 PHE200 GLU201 LYS202 CYS309 GLU375 MET378 GLN404 LEU405 PRO406 ALA407 PRO408 VAL484 SER486 PRO514 TYR515 ASN516 SER517 ASN519 THR532 VAL533 ASP534 SER535 SER536 GLN537 ARG560 VAL563
	5RML	SiteA2	Lys139, Glu142, Glu143, Lys146, Tyr149, Asn179, Tyr180, Pro408, Thr410, water	TRP114 TYR120 ASN124 LEU138 LYS139 THR141 GLU142 PHE145 LYS146 TYR149 ASN179 TYR180 THR380 ASN381 TYR382 ASP383 PRO408 ARG409 THR410 LEU411 LEU412 LEU417 TYR421 PHE422 ASN423

Table 3.6. The AA residues present at the two sites (A1 and A2) of SC-2 NSP13 along with the residues predicted by SiteFinder. The binding site was described to be within 4.5 Å. The site finder resulted in the amino acids and their positions for the two sites.

The dummy atoms were present within the fragment binding pockets. Most of the amino acid residues present within these binding pockets were found to be conserved in SC-2 NSP13 variants (GISAD dataset-green) as well as in betacoronaviruses (blue) (Figure 3.18.). The unhighlighted ones were within the intermediate conservation threshold.

(a)

LEU176 ASN177 ARG178 ASN179 PHE200 GLU201 LYS202 CYS309 GLU375 MET378 GLN404 LEU405
PRO406 ALA407 PRO408 VAL484 SER486 PRO514 TYR515 ASN516 SER517 ASN519 THR532 VAL533
ASP534 SER535 SER536 GLN537 ARG560 VAL563

LEU176 ASN177 ARG178 ASN179 PHE200 GLU201 LYS202 CYS309 GLU375 MET378 GLN404 LEU405
PRO406 ALA407 PRO408 VAL484 SER486 PRO514 TYR515 ASN516 SER517 ASN519 THR532 VAL533
ASP534 SER535 SER536 GLN537 ARG560 VAL563

(b)

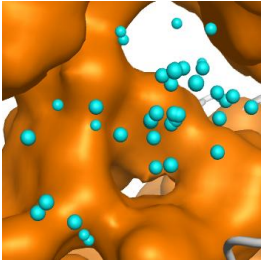
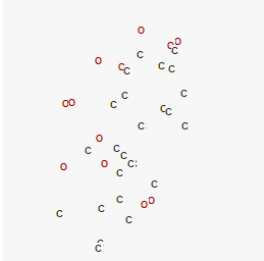
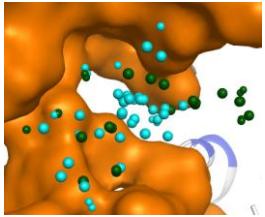
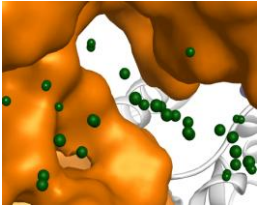
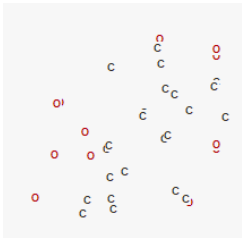
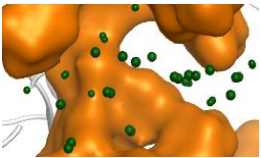
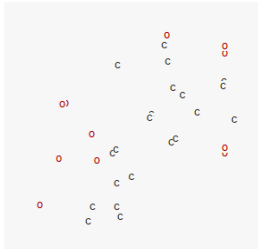
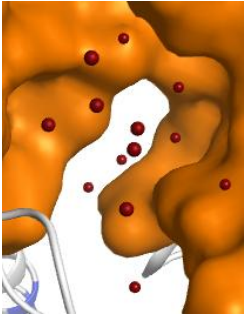
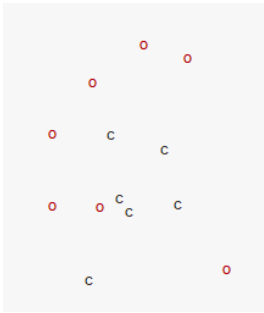
TRP114 TYR120 ASN124 LEU138 LYS139 THR141 GLU142 PHE145 LYS146 TYR149 ASN179 TYR180
THR380 ASN381 TYR382 ASP383 PRO408 ARG409 THR410 LEU411 LEU412 LEU417 TYR421 PHE422
ASN423

TRP114 TYR120 ASN124 LEU138 LYS139 THR141 GLU142 PHE145 LYS146 TYR149 ASN179 TYR180
THR380 ASN381 TYR382 ASP383 PRO408 ARG409 THR410 LEU411 LEU412 LEU417 TYR421 PHE422
ASN423

Figure 3.18. (a) SiteA1 binding site amino acid residues with highlighted conserved amino acids. (b) SiteA2 binding site amino acid residues with highlighted conserved amino acids. The conserved residues from the GISAID dataset are highlighted in green whereas the conserved residues from the betacoronaviruses dataset are highlighted in blue.

Virtual screening (VS) workflow

After the SiteFinder dummy atom placements, the two sites were sub-divided into smaller pockets. SiteA1 was divided into three small pockets and one large pocket, whereas site A2 was divided into five small pockets and one large pocket as well (Table 3.7.). Each pocket was colour coded and sdf files were generated for fastROCS screening. In the case of siteA1 there were two pockets (pocket 2 and pocket 2.1) representing a similar site but with slight difference in the cavity point placements. Similarly, siteA2 also had three pockets (pocket 1, pocket 4 and pocket 5) representing the same target site.

3 Fragment (5RLH, 5RMM, 5RLZ)	Dummy atoms in 3D	Dummy atoms in 2D	All Dummy atoms together
Pocket 1			
Pocket 2			
Pocket 2.1			
Large volume dummy atom for 3 fragments site			

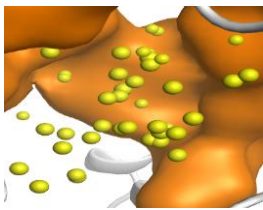
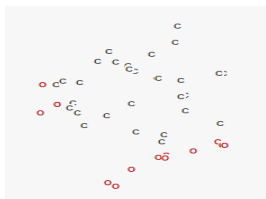
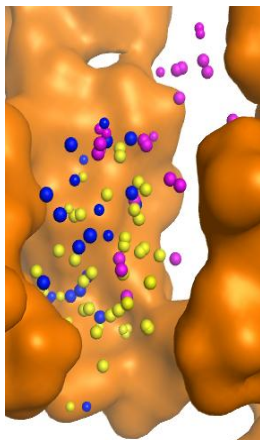
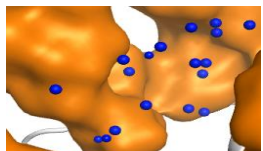
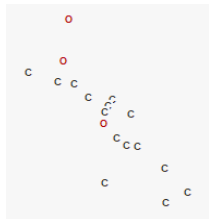
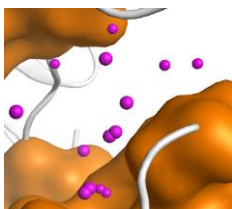
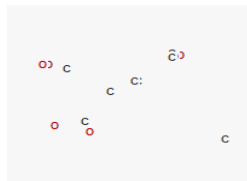
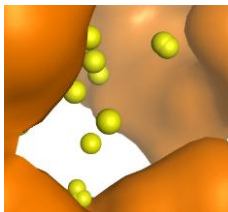
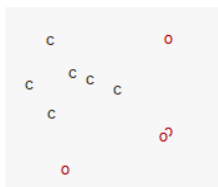
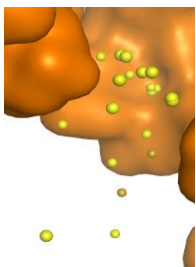
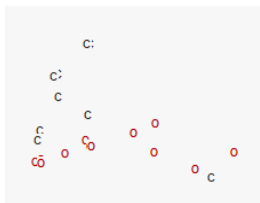
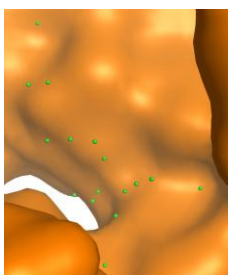
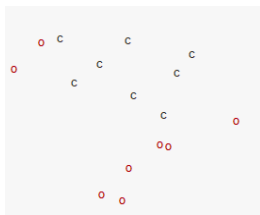
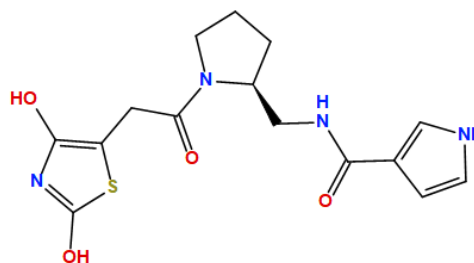
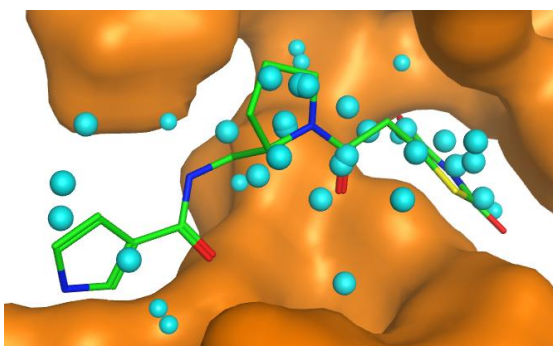
1 Fragment (5RML)	Dummy atoms in 3D	Dummy atoms in 2D	All Dummy atoms together
Pocket 1			
Pocket 2			
Pocket 3			
Pocket 4 Extension of pocket 1			
Pocket 5 below pocket 1			
Large volume dummy atom for 1 fragment site			

Table 3.7. Different pockets of sites A1 and A2 with dummy atom placements on the RNA-binding tunnel of SC-2 NSP13. The dummy atoms are depicted in both 3D and 2D formats. In the 2D representation, the oxygen atoms are coloured red to indicate hydrophilic cavity points, while the carbon atoms are shown in black to represent hydrophobic cavity points.

1. The first round of screening was **fastROCS** volume query search on the Enamine database. Hits from the screening for each pocket mapped well with the dummy atom query (Figure 3.19.). The fastROCS screening resulted in 2,400,000 hits for siteA1 whereas, for siteA2 3,600,000 hits were predicted.

(a)



(b)

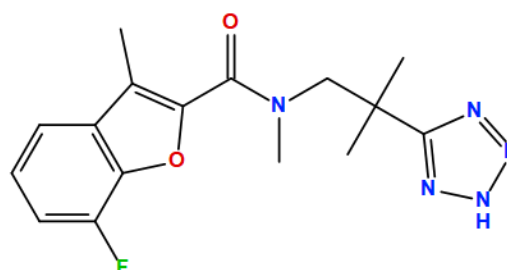
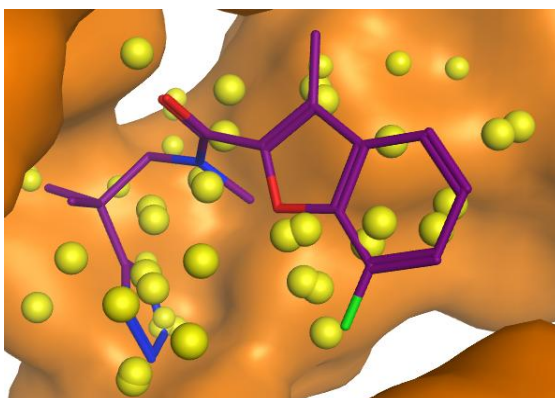
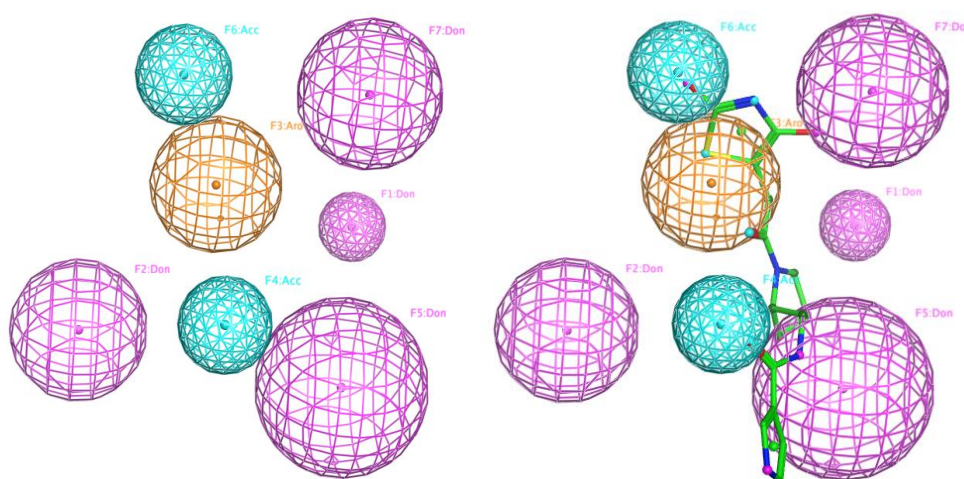


Figure 3.19. (a) One of the hits PV-004327695585 (green) of pocket 1 (siteA1-NSP13) mapping well with the dummy atoms (cyan). (b) One of the hits Z4444958600 (purple) of pocket 1 (siteA2-NSP13) mapping well with the dummy atoms (yellow).

2. The hits from the fastROCS screening were the input into the next round of VS i.e., **pharmacophore** search. MoPBS generated a seven features pharmacophore for siteA1 whereas, for siteA2 a five features pharmacophore query was generated. In the case of

siteA1 a partial search was done with at least five features covered. This was done as there were no hits that had all the seven features. The output of the pharmacophore search resulted in 9,165 hits for siteA1 and 142,162 hits for siteA2. The hits mapped well with the pharmacophore query (Figure 3.20.). The hit lists were filtered using MOE database viewer which resulted in 8,906 hits for siteA1 and 141,343 for site A2, after the removal of molecules having PAINS, molecular weight > 550 g/mol, and SlogP > 5.

(a)



(b)

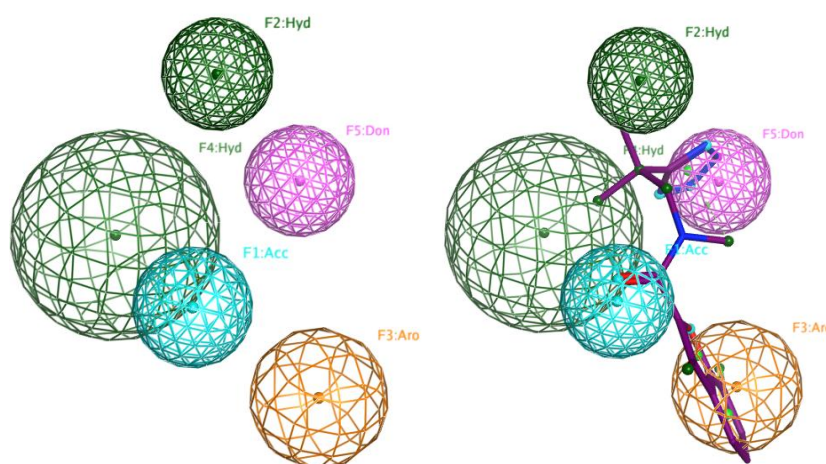
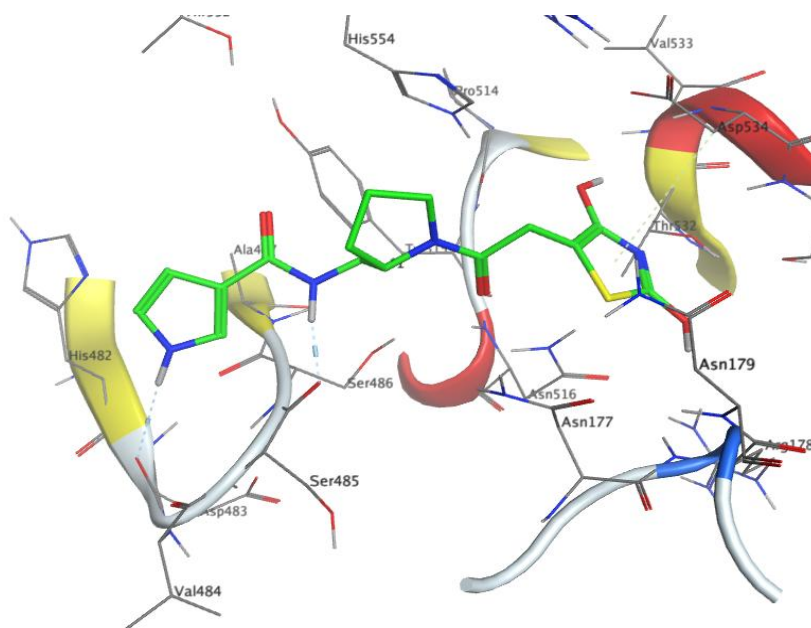


Figure 3.20. (a) The pharmacophore feature query for siteA1 of NSP13 and one of the hits PV-004327695585 (green) from the search mapping with the features (b) The pharmacophore feature query for siteA2 of NSP13 and one of the hits Z4444958600 (purple) from the search mapping with the features. Hyd (dark green)- Hydrophobic

feature, Don (pink)- Hydrogen bond donor feature, Aro (orange)- Aromatic ring feature, Acc (cyan)- Hydrogen bond acceptor feature.

- The last round of VS was **molecular docking**. The hit lists after filtering were used for docking at siteA1 and siteA2 (Figure 3.21.). The output of docking resulted in 7,687 hits for siteA1 and 127,961 hits for siteA2. The molecules from these hit lists were then clustered together to form a diverse set of 150 representative molecules using Pipeline Pilot. The FRED docking score (Chemgauss4 score) was in the range of -11.71 to -6.55 of these 150 hits.

(a)



(b)

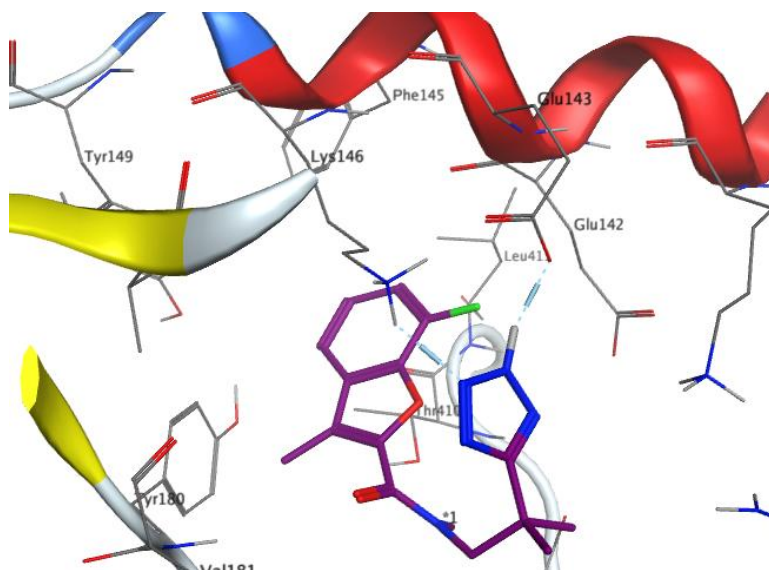


Figure 3.21. (a) Docking of one of hits PV-004327695585 (green) at siteA1 of NSP13 making interactions with conserved AA residues Asp534, Val484 and Ser485. (b) Docking of one of hits Z4444958600 (purple) at siteA2 of NSP13 making interactions with highly conserved AA residues Lys146 and Glu143.

The 150 hits at the end of VS fulfilled both the fastROCS volume and pharmacophore features queries. The best interaction pose between the hit and protein was obtained by docking. The hits were observed to make interactions with AA residues that were conserved across the NCBI, GISAID and the betacoronaviruses datasets. Upon visual inspection 92 final hits were sent to CACHE for experimental validation. The Figure 3.22. summarises the entire tiered VS workflow of the MDG group.

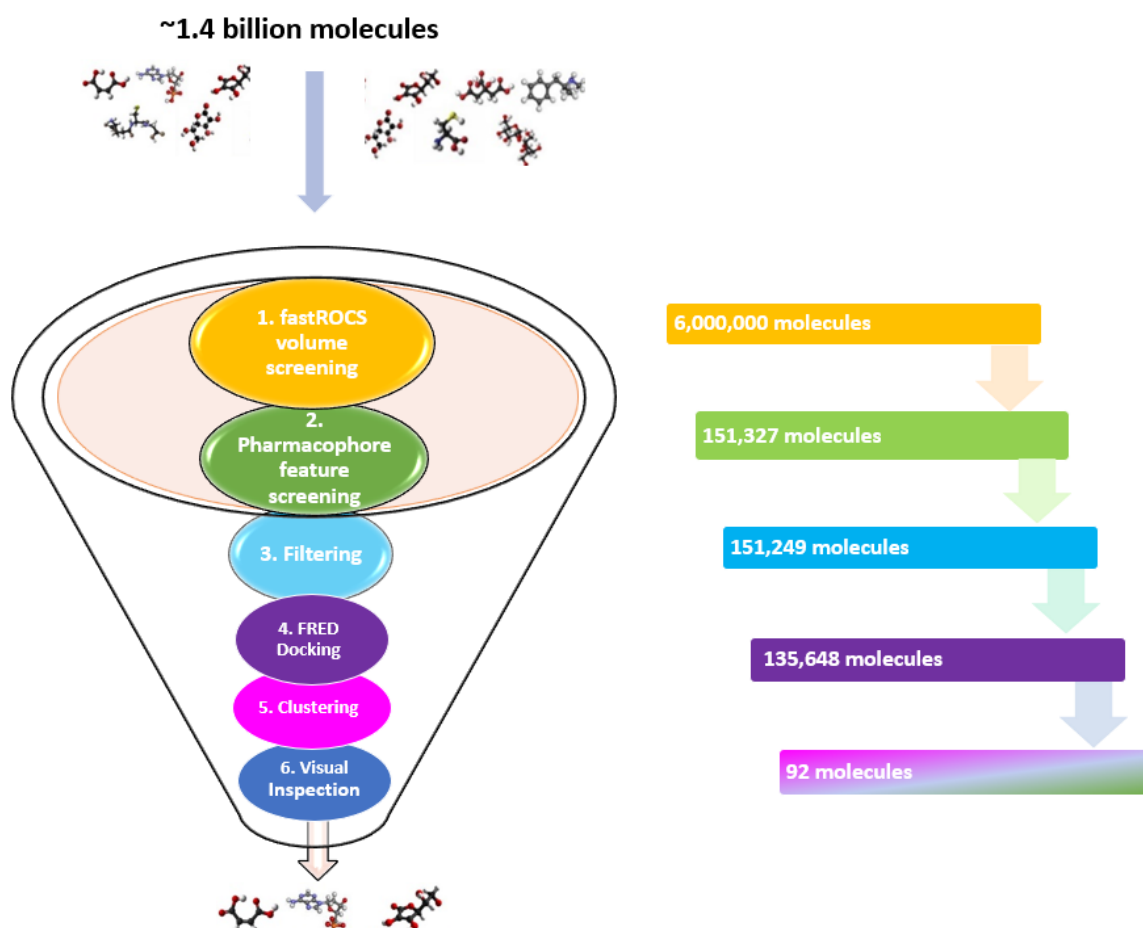


Figure 3.22. MDG tiered virtual screening workflow for SC-2 NSP13. The workflow involved screening a database of approximately 1.4 billion molecules using fastROCS, followed by pharmacophore analysis, filtering, FRED docking, clustering, and ultimately visual inspection, which led to the identification of 92 molecules.

(C) Experimental validation results by the CACHE organisation

The CACHE organisation received submissions for predicted hits from 23 participants for challenge #2. The experimental validation was conducted in two rounds. For round 1, predicted hits from the submission (1,957 compounds) were screened at 50 μ M against SC-2 NSP13 in a Surface Plasmon Resonance (SPR) binding assay. The hit candidates were also tested in a dose-response experiment with different concentrations of the compounds and the NSP13 ATPase assay. In this round, the aggregation and solubility of the compounds were also measured by dynamic light scattering (DLS). The compounds of interest from round 1 were then advanced to the hit-expansion phase, i.e., round 2, wherein 50 analogues were sourced for biological validation.

In round 1, we had submitted 92 compounds for experimental testing. The compounds were screened in a primary SPR binding assay at 50 μ M against SC-2 NSP13. The data from this assay is the ratio of the measured SPR signal in response units, or RU, over the maximal expected signal. The compounds with RU ratios between 0.5 and 2.0 (**CACHE cut-off**) were selected for dose-response experiments; the compounds with RU > 2.0 were excluded as they are indicative of non-specific binding. In this round, five hits from our submission with RU in the range of 0.49–1.09 were selected for the dose-response experiment and NSP13 ATPase assay (Table 3.8.) (Figure 3.23.).

CACHE ID	SMILES	RU (50uM)
CACHE_1432_22	<chem>C[C@H](C(C)(C)c1nn[nH]n1)NC(c1cc(cc2cccnc12)F)=O</chem>	0.94
CACHE_1432_28	<chem>C1[C@H]2CN(Cc3cccnc3N2C[C@@H]1O)c1c2cn[nH]c2ncn1</chem>	1.09
CACHE_1432_51	<chem>C1CC(CCC1CNC(CCC1=CC=CC(N1)=O)=O)c1nn[nH]n1</chem>	0.64
CACHE_1432_77	<chem>Cc1c(cc2cc(ccc2n1)c1ccc2c(c1)c(nc(n2)O)O)O</chem>	0.17
CACHE_1432_88	<chem>C1CN(Cc2c1nc[nH]2)C(c1cc(c2ccncc2)nnc1O)=O</chem>	0.49

Table 3.8. List of CACHE hits that passed in primary SPR binding assay at 50 μ M against SC-2 NSP13. The CACHE IDs of compounds, the SMILES strings and their RU values are mentioned. These results were produced by the CACHE organisation.

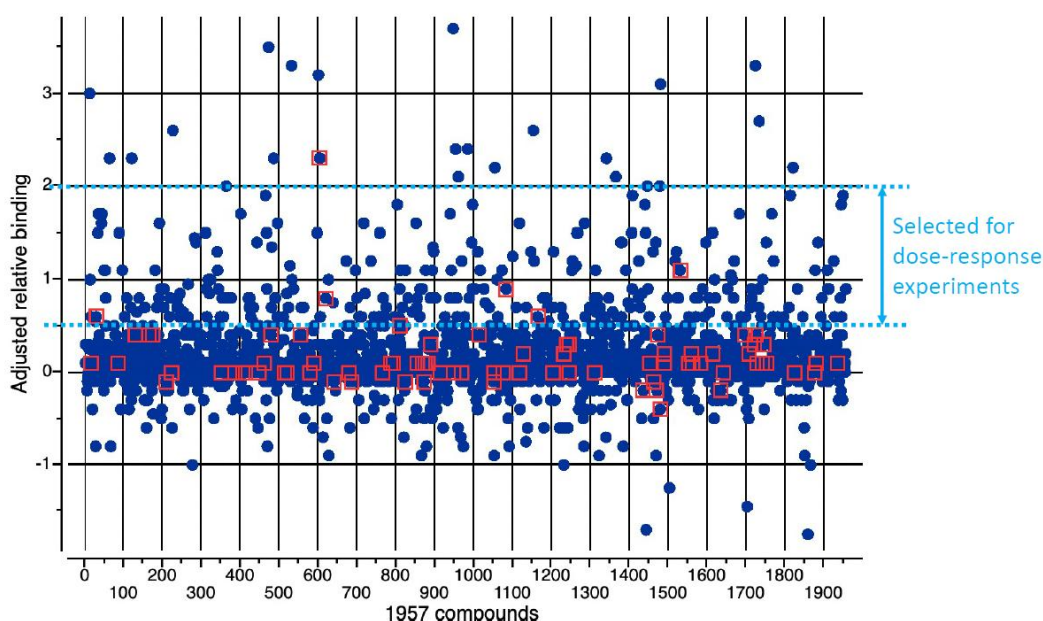


Figure 3.23. CACHE hits (red boxes) that passed in primary SPR binding assay at 50 μ M against SC-2 NSP13. Only the hits with RU ratios between 0.5 and 2.0 (blue dotted lines) were selected for further testing. The y-axis shows adjusted relative binding (RU) and the x-axis shows test compounds. Red boxes highlight the compounds chosen for this study's dose-response experiments between 0.5 and 2.0 RU values. These results were produced by the CACHE organisation.

In the dose-response experiments, the compounds were tested against NSP13 with duplicate runs. Only the CACHE_1432_28 compound showed acceptable binding with K_d 551 μ M with 268% binding (stronger RU signal) in the first run though % binding > 150 indicates non-specific binding. But in the second replicate none of the hits showed binding. These compounds were also subjected to the NSP13 ATPase inhibition assay, which measures how much the compound inhibits NSP13's ATP hydrolysis activity. It was found that only compound CACHE_1432_77 showed 82%-102% ATPase inhibition even though it did not show any binding in the SPR assay. The inhibition could be a result of allosteric binding (Table 3.9.). The compound CACHE_1432_88 was not included in the aggregation and solubility testing as it showed no binding in the SPR and no inhibition in the ATPase assay.

	SPR binding assay: NSP13 dose response				NSP13 Catalytic ATPase assay	
CACHE ID	NSP13 K_d (μ M) [1]	% Binding [1]	NSP13 K_d (μ M) [2]	% Binding [2]	ATPase % inhibition at 50 μ M [1]	ATPase % inhibition at 50 μ M [2]
CACHE_1432_22	>1000	>1000	0	0	10	21
CACHE_1432_28	551	268	0	0	11	22
CACHE_1432_51	0	0	0	0	0	7
CACHE_1432_77	>1000	>1000	0	0	82	102
CACHE_1432_88	>1000	>1000	0	0	17	ND

Table 3.9. List of hits with dose response SPR binding assay results along with ATPase assay results. ND-not determined, [1] first run, [2] second run. These results were produced by the CACHE organisation.

To check if the compounds have good solubility and no aggregation/clogging at the concentration they are found active, the DLS method is used. It measures the amount of aggregation in kilo-counts per second (kcnt/s) by measuring the intensity of the scattered light. The scattering of light was observed at different concentrations of the compounds. The intensity of the scattered light will be high if there are larger aggregates (O'Donnell et al., 2021). Only compound CACHE_1432_77 showed acceptable aggregation at all concentrations. Compound CACHE_1432_51 demonstrated elevated intensities across all concentrations, suggesting the presence of larger aggregates. For the solubility determination the % laser power is used to determine scattering of light. A lower % laser power means more scattering due to high particulate matter and thus will have poor solubility. Only compound CACHE_1432_51 had poor solubility compared to the rest of the compounds, which had good solubility (Table 3.10.). Unfortunately, none of the hits were advanced to round 2 for analogue testing because, for the dose response SPR binding assay, the CACHE organisation set the cut-off to $K_d < 150 \mu\text{M}$. The steps involved in the experimental validation performed by the CACHE organisation is depicted in Figure 3.24. The five hits from round 1 experimental validation results are shown in (Table 3.11.) with docked score and site where they were docked.

CACHE ID	DLS aggregation assay: Intensity (kcnt/s)				DLS solubility assay: % laser power		
	200 μ M	100 μ M	50 μ M	25 μ M	100 μ M	50 μ M	25 μ M
CACHE_1432_22	1693	615	505	496	100	100	100
CACHE_1432_28	1552	423	471	547	100	100	100
CACHE_1432_51	4420	3298	4349	3294	35	45	85
CACHE_1432_77	290	1043	448	607	100	100	100

Table 3.10. List of hits with aggregation and solubility results. DLS aggregation assay results expressed as scattering intensity (kcnt/s) at varying compound concentrations (200, 100, 50, and 25 μ M), and DLS solubility assay results expressed as percent laser power (% laser power) at 100, 50, and 25 μ M concentrations. These results were produced by the CACHE organisation.

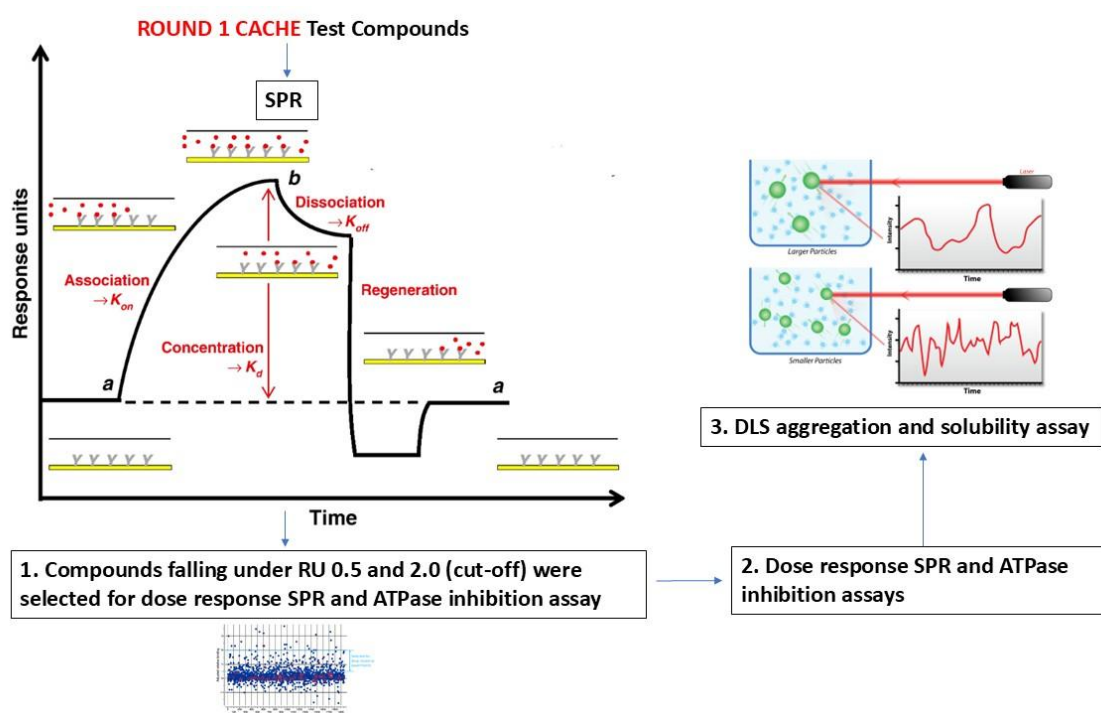
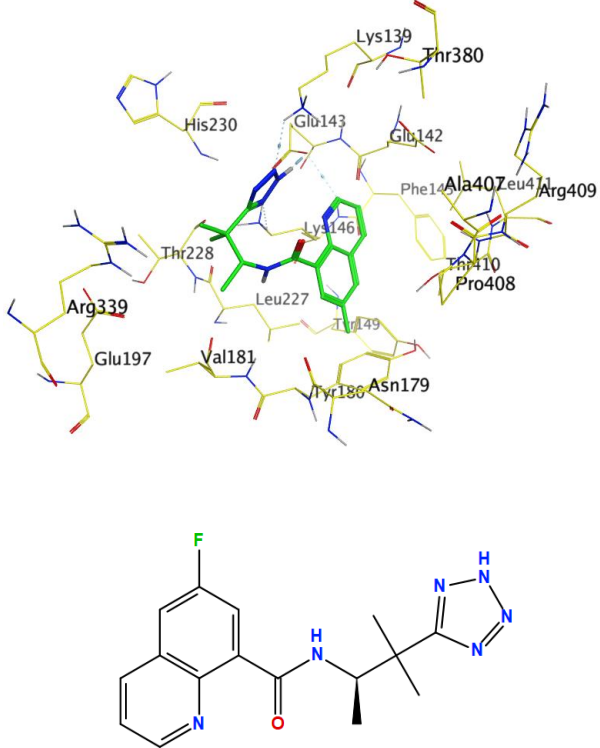
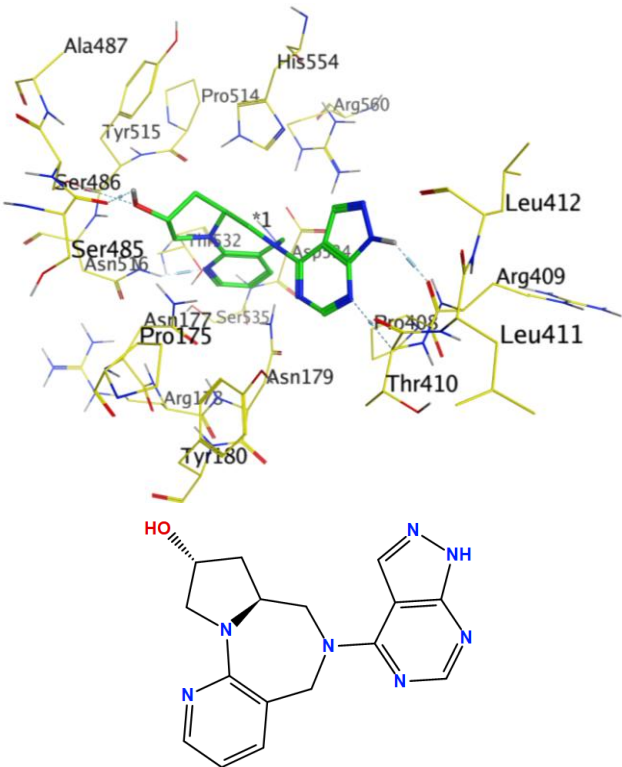
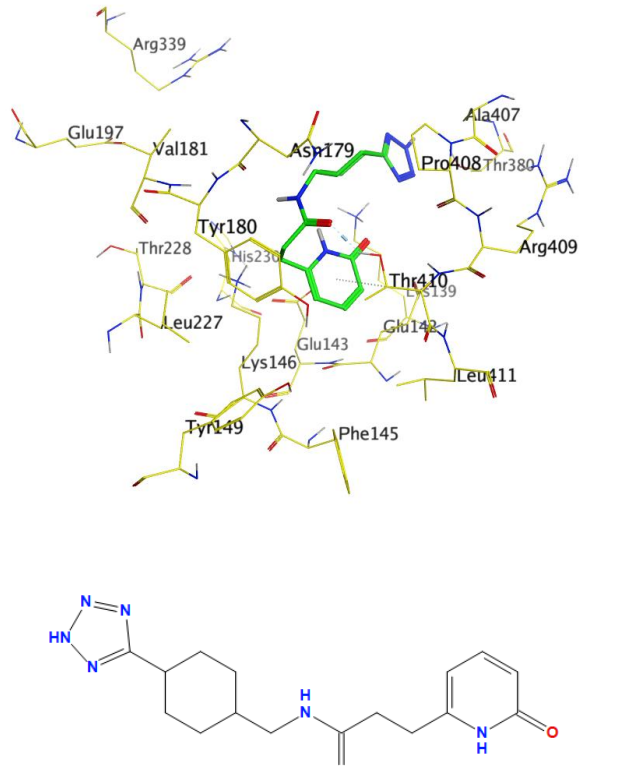


Figure 3.24. The experimental validation performed by the CACHE organisation involved various steps. (1) The test compounds were subjected to an SPR binding assay, which resulted in a sensorgram. The x-axis shows time and the y-axis shows (RU). The sensorgram shows that initially the protein (grey) is immobilised on the chip, and upon injecting the test molecules (red), if there is interaction between the test molecule and the protein, the refractive index at the surface of the chip changes, and the consequence is seen as an increase in signal intensity. The increase in the signal is described by the resonance or response units (RU). The initial binding phase is known as the association

phase (Kon). Upon achieving a steady state, the maximum RU relates to the concentrations of immobilised protein and test molecules and is thus used for measuring binding affinity (Kd). When the test molecules are removed due to continuous flow, there is a dissociation phase (Koff). The experiment can later be restarted during the regeneration phase. The schematic was taken from (Patching, 2014). If interaction between the immobilised receptor molecule and the analyte molecule occurs, the refractive index at the surface of the gold film changes, and the outcome is seen as an increase in signal intensity. (2) The test molecules that passed the first round of SPR were then subjected to dose-response SPR and ATPase inhibition assays. (3) Later, the selected molecules from step (2) were subjected to a DLS solubility and aggregation assay (schematic taken from Wikipedia DLS - Dynamic light scattering - Wikipedia). If the aggregates are larger (green), the intensity of the scattered light will increase and vice versa. In solubility determination, lower percentage laser power indicates increased scattering from high particulate matter, resulting in poor solubility, and vice versa.

Compound	Docked pose	Docking score	Site
CACHE_1432_22		-9.9	SiteA2

CACHE_1432_28		-9.7	SiteA1
CACHE_1432_51		-9.2	SiteA2

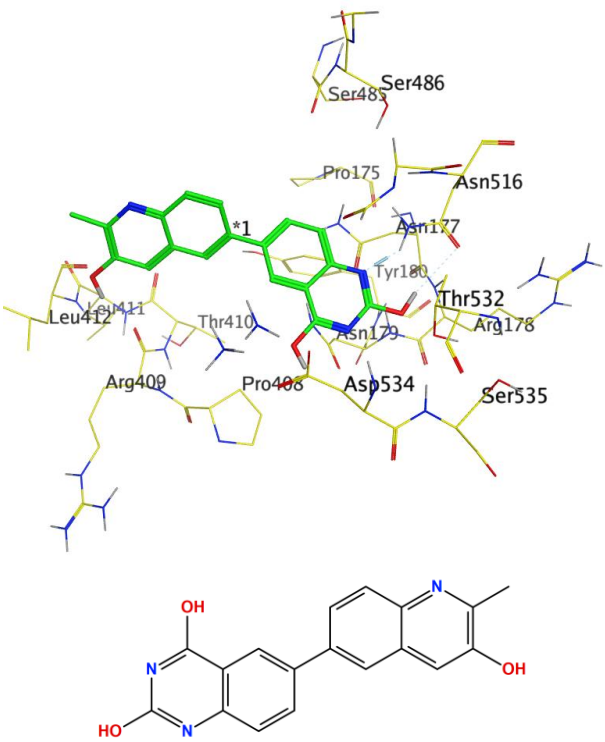
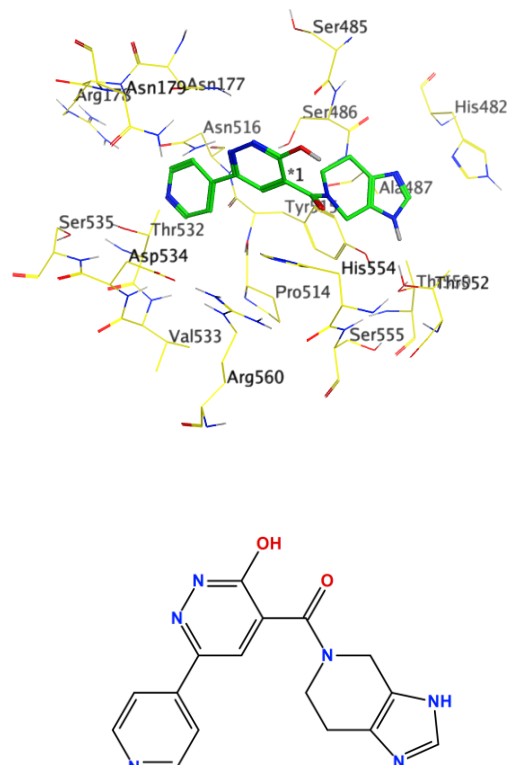
CACHE_1432_77		-8.2	SiteA1
CACHE_1432_88		-8.5	SiteA1

Table 3.11. Five hits (green) from round 1 experimental validation results with docked poses, docking score (kcal/mol) and site of docking (yellow). The dotted lines represent hydrogen bonds.

Protocol validation

The CACHE organisation released the list of compounds (actives) that showed binding to NSP13 in round 1 and round 2 of experimental validation. To test and improve our VS protocol, we performed the entire workflow on these actives, as well as the compounds that did not show binding to NSP13 as inactive. For round 1, there were 46 actives and 1911 inactives, and for round 2, there were 45 actives and 505 inactives.

When we used our VS workflow on these datasets, it was found that for round 1 all the actives were found as hits in fastROCS VS for both siteA1 and siteA2. However, the pharmacophore VS only identified one active as a hit for siteA1 but none for siteA2. Our pharmacophore VS stage was too stringent at filtering potential actives. The pharmacophore feature query was changed from a partial search with at least 5 features to at least 4 features for siteA1 and from all 5 features to a partial search with at least 4 features for siteA2. These improvements of the pharmacophore features helped in identifying more actives in the pharmacophore VS stage. For siteA1, 23 actives were identified, and for siteA2 38 actives were identified.

In the case of round 2, when we used our VS workflow, it was found that all the actives were found as hits in fastROCS VS for both siteA1 and siteA2. Using our pharmacophore VS, only 13 actives were identified as hits for siteA1, and 24 actives were identified as hits for siteA2. Again, it was the pharmacophore query, that was too tightly defined. The pharmacophore feature query was changed from partial search with at least 5 features to at least 4 features for siteA1 and from all 5 features to partial search with at least 4 features for siteA2. These improvements of the pharmacophore features helped in identifying more actives in the pharmacophore VS stage. For siteA1, 20 actives were identified, and for siteA2 35 actives were identified. The validation protocol thus helped modify the VS workflow for use in future VS campaigns.

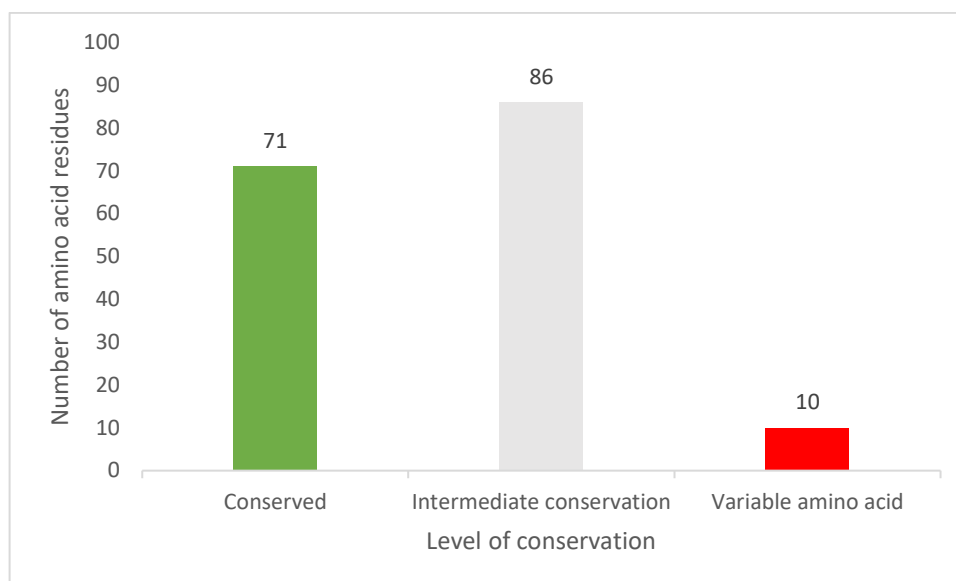
3.7.2. NSP3 results

(A) NSP3 bioinformatics analysis

From the NCBI dataset analysis on **1,070,751** SC-2 Mac1 NSP3 sequences (167 AA in length) it was found that there were 71 conserved amino acids, 86 intermediate conservation and top 10 highly variable amino acids (Figure 3.25.) which are discussed in section 2.9.9. of

Chapter II. The mutational and conservational threshold selected were 0.171% and 0.013% respectively.

(a)



(b)

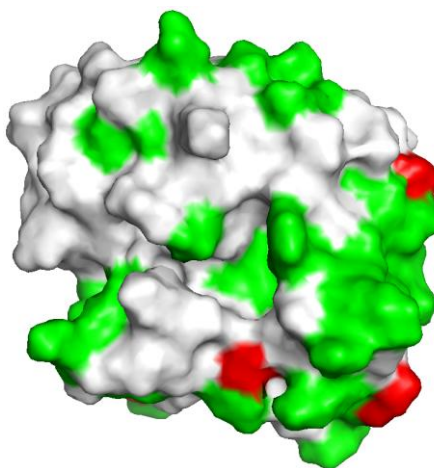


Figure 3.25. (a) Bar graph with number of amino acids belonging to conserved (green), intermediate conservation (white) and variable amino acid (red). The x-axis shows the level of conservation, and the y-axis shows number of amino acid residues. (b) Visualisation of these residues on a 3D protein structure using MOE using same colour scheme as in (a).

As per the mutational analysis on the NCBI dataset using BioAider, it was found that among the top 10 mutations, there were a few mutations that caused changes in amino acid properties. The mutation S164L (SER164LEU) led to a change from being polar and uncharged to non-polar and uncharged. The mutation D103Y (ASP103TYR) led to a change

from being polar, negative-charged to polar, uncharged. The mutation T31A (THR31ALA) led to a change from being polar, uncharged to non-polar, uncharged. The mutation K74N (LYS74ASN) led to a change from being polar, positive-charged to polar, uncharged. The mutation T144I (THR144ILE) led to a change from being polar, uncharged to non-polar, uncharged. And lastly, the mutation P134S (PRO134SER) led to a change from being non-polar, uncharged to polar and uncharged (Figure 3.26.).

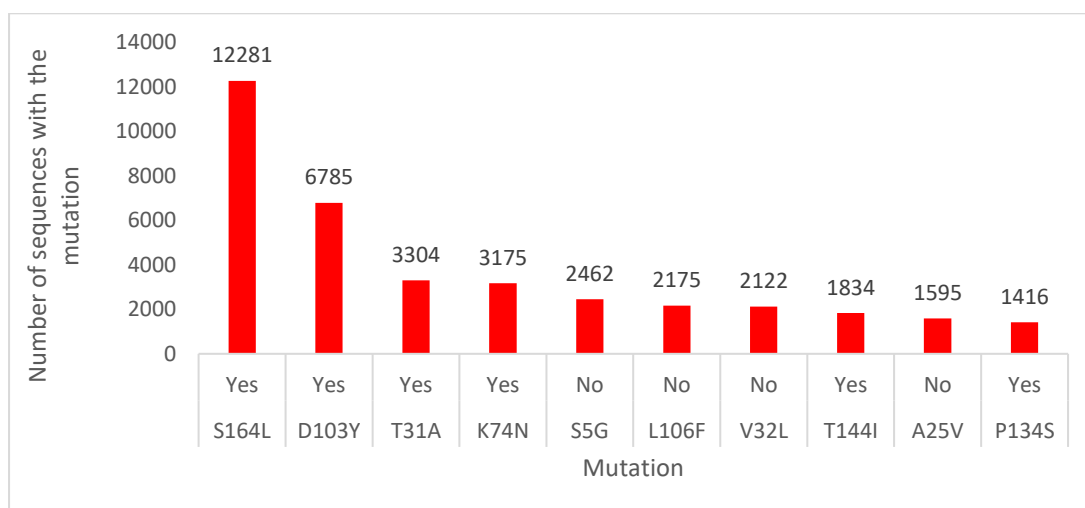
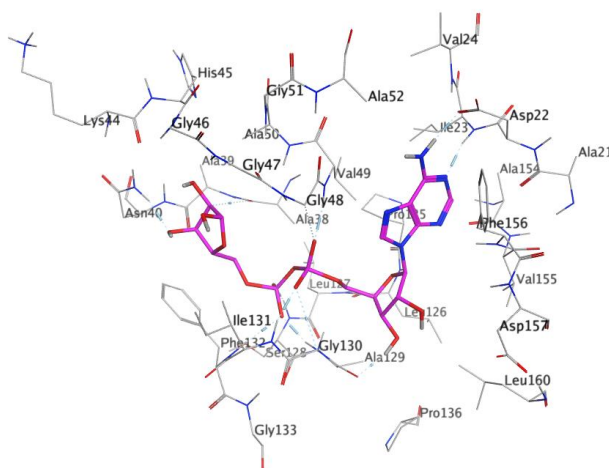


Figure 3.26. Bar graph showing the top 10 mutations from the NCBI SC-2 Mac1 domain of NSP3 sequence dataset. Yes- mutation changed the amino acid properties, No- mutation did not change the amino acid properties. The x-axis shows the type of mutation, and the y-axis shows the number of sequences with the mutation.

In this study, the ADP-r binding site (PDB: 7KQP) and the site where the lead-like compounds were co-crystallised (Gahbauer et al., 2023) were targeted (Figure 3.27.). The amino acid residues present in the binding pockets were found to be within the conserved and intermediate conservation thresholds in the NCBI dataset (Table 3.12.). These binding sites were targeted for the identification of small molecule inhibitors in **challenge #3**.

(a)



(b)

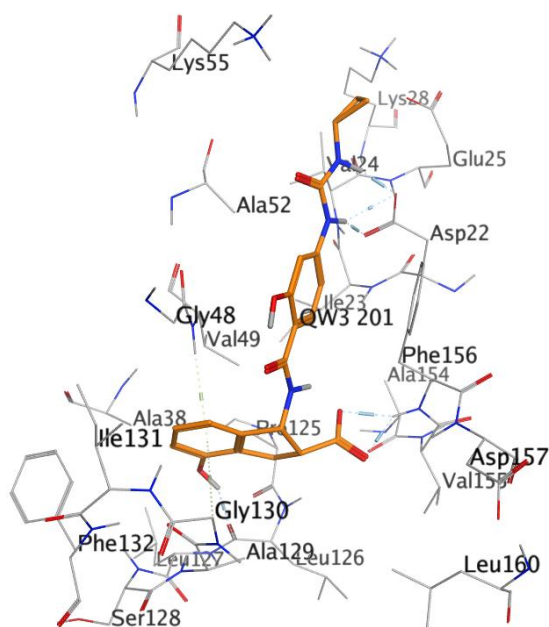


Figure 3.27. (a) ADP-r (pink) binding site of SC-2 Mac1 domain of NSP3 (PDB: 7KQP). (b) Crystal structure of a lead-like compound (orange) co-crystallised in the Mac1 domain of NSP3 (PDB: 5SQW).

AA position	Amino acid	Number of sequences	Conservation percent	gaps	Mutations	Mutational threshold
19	A	1070017	99.93145	0	V (495); T (156); D (49); S (22); E (12)	0.069
20	D	1070399	99.96713	0	Y (224); N (95); A (13); G (12); E (7); K (1)	0.033
21	I	1070577	99.98375	0	V (165); L (5); T (2); F,M (1)	0.016

22	V	1070605	99.98636	0	M (101); A (38); L (4); G (3)	0.014
23	E	1070483	99.97497	0	K (137); G (69); D (30); A (16); I (9); V (4); Q (2); R (1)	0.025
26	K	1070718	99.99692	0	E (22); N,R (5); T (1)	0.003
36	A	1070547	99.98095	0	V (166); T (32); S (6)	0.019
37	A	1070562	99.98235	0	T (90); V (88); S (8); P,H,D (1)	0.018
38	N	1070570	99.9831	0	S (142); T (30); D (4); H (3); I (2)	0.017
42	K	1070326	99.96031	0	R (249); E (172); N (4)	0.040
43	H	1070554	99.9816	0	Y (153); Q (35); N (6); R,C,D (1)	0.018
44	G	1070490	99.97562	0	E (122); R (121); V (18)	0.024
45	G	1070502	99.97675	0	R (169); E (65); V (15)	0.023
46	G	1070608	99.98664	0	C (78); S (54); V (8); D (3)	0.013
47	V	1069672	99.89923	0	I (841); A (216); F (19); G (3)	0.101
48	A	1070716	99.99673	0	T (12); V (11); S (10); G,P (1)	0.003
49	G	1070085	99.9378	0	V (332); R (241); E (89); A (4)	0.062
50	A	1069558	99.88858	0	V (852); T (233); S (78); D (30)	0.111
53	K	1070604	99.98627	0	R (85); N (36); E (19); T (4); M (3)	0.014
123	P	1070643	99.98991	0	S (98); T (8); L,I (1)	0.010
124	L	1070737	99.99869	4	V,I (4); F (2)	0.001
125	L	1070741	99.99907	4	V (3); I,Y,F (1)	0.001
126	S	1070584	99.9844	1	A (67); L (64); P (34); M (1)	0.016
127	A	1070182	99.94686	0	V (488); T (36); S (35); D (9); P (1)	0.053
128	G	1070448	99.9717	6	S (154); D (127); C (11); V (5)	0.028
129	I	1070249	99.95312	0	V (481); T (14); F (5); S,L (1)	0.047
130	F	1070722	99.99729	1	C (13); V (9); L (4); Y,S (1)	0.003
131	G	1069567	99.88942	0	C (1029); D (82); S (50); V (22); F (1)	0.111
134	P	1067895	99.73327	0	S (1416); L (1412); T (15); H (12); F (1)	0.267
152	A	1070271	99.95517	0	V (401); T (41); G (16); S (13); D (7); P,L (1)	0.045
153	V	1070607	99.98655	0	I (94); F (37); G (7); A (5); S (1)	0.013
154	F	1070726	99.99767	0	S (15); V,L (4); I,C (1)	0.002

155	D	1070570	99.9831	0	G (70); N (37); E (30); Y (29); A (14); I (1)	0.017
158	L	1070062	99.93565	0	F (604); I (30); R (27); P (22); V,H (3)	0.064

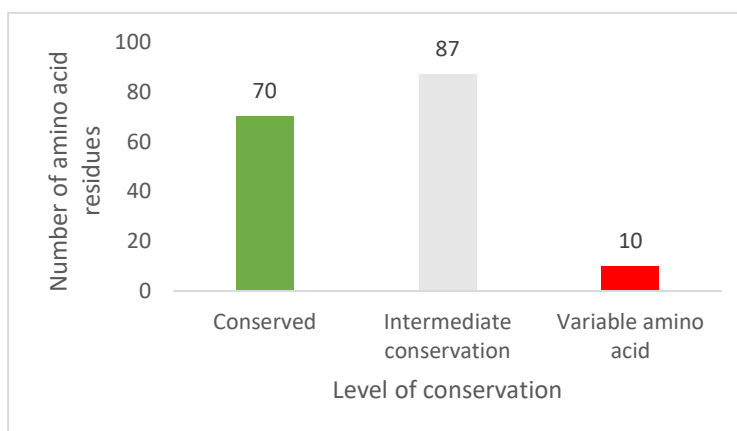
Table 3.12. List of amino acid residues present in the binding site of SC-2 Mac1 domain of NSP3. The conserved residues are highlighted in green, and the intermediate conservation amino acids are highlighted in yellow (NCBI dataset). The AA position column shows the position of amino acid within the Mac1 domain NSP3 sequence, amino acid column shows the type of amino acid, number of sequences are the sequences having that exact amino acid, the conservation percent is the percentage of sequences having to amino acid as the reference sequence, gaps column shows if any gaps were inserted, mutations column shows the type of amino acid change along with the number of sequences having that mutated amino acid in brackets and mutational threshold column which shows the percentage of mutation.

From the GISAID dataset analysis on **7,104,871** SC-2 Mac1 NSP3 sequences (167 AA in length), it was found that there were 70 conserved amino acids, 87 intermediate conservation and top 10 highly variable amino acids (Figure 3.27. a). The mutational and conservational thresholds selected were 0.178% and 0.026%, respectively. The top mutations were SER164LEU, ILE67VAL, HIS89TYR, GLY101CYS, PRO134LEU, THR31ALA, THR144ILE, ASP103TYR, LEU158PHE and VAL119PHE. Among them, mutations ILE67VAL, HIS89TYR, GLY101CYS, LEU158PHE and VAL119PHE were not identified as top 10 mutations in the NCBI dataset. This result points out the importance of sequence dataset size in the identification of populated mutations. The two datasets (NCBI and GISAID) also shared common mutations SER164LEU, THR31ALA, THR144ILE, ASP103TYR and PRO134 position mutated to two different amino acids, LEU (GISAID) and SER (NCBI). These mutations are populated across the SC-2 variants and might hold biological significance.

As per the mutational analysis on the GISAID dataset using BioAider, it was found that among the top 10 mutations, there were a few mutations that caused changes in amino acid properties. The mutation S164L (SER164LEU) led to a change from being polar and uncharged to non-polar and uncharged. The mutation D103Y (ASP103TYR) led to a change from being polar, negative-charged to polar, uncharged. The mutation T31A (THR31ALA) led to a change from being polar, uncharged to non-polar, uncharged. The mutation H89Y (HIS89TYR) led to a change from being polar, positive-charged to polar, uncharged. And

lastly, mutation T144I (THR144ILE) led to a change from being polar, uncharged to non-polar, uncharged (Figure 3.28. b).

(a)



(b)

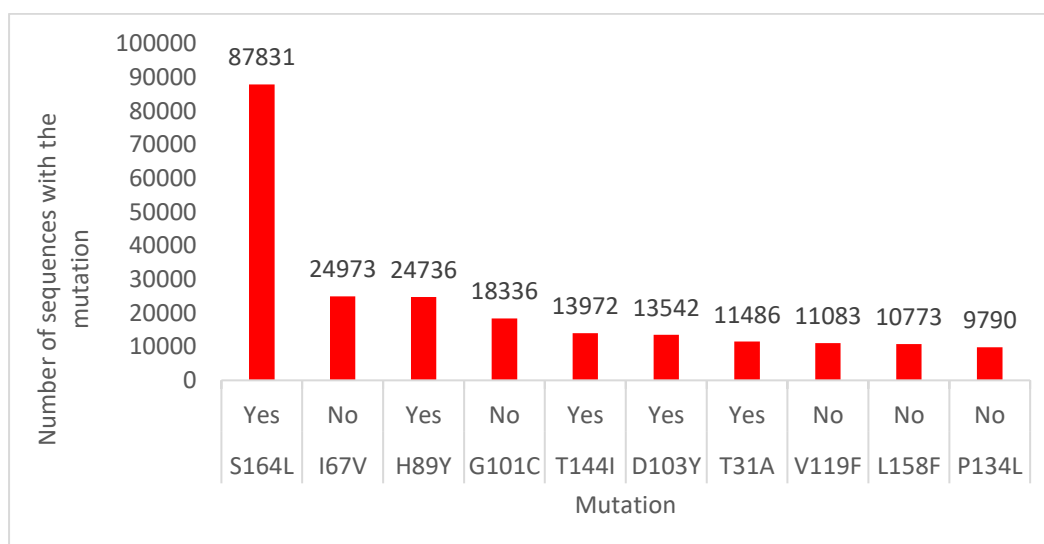


Figure 3.28. (a) Bar graph with number of amino acids belonging to conserved (green), intermediate conservation (white) and variable amino acid (red). The x-axis shows the level of conservation, and the y-axis shows number of amino acid residues. **(b)** Bar graph showing the top 10 mutations from the GISAID SC-2 Mac1 NSP3 sequence dataset. Yes- mutation changed the amino acid properties, No- mutation did not change the amino acid properties. The x-axis shows the type of mutation, and the y-axis shows the number of sequences with the mutation.

AA position	Amino acid	Number of sequences	Conservation percent	gaps	Mutations	Mutational threshold
19	A	7099774	99.92826	0	V (3118); N (1014); T (713);	0.0717

					S (117); D (67); E (51); G (2)	
20	D	7102096	99.96094	0	Y (1104); A (1068); N (374); G (127); E (65); H (21); I (1)	0.0391
21	I	7103015	99.97388	0	D (1014); V (782); L (28); F (13); T (3); S (1)	0.0261
22	V	7100832	99.94315	0	M (2816); I (1015); A (133); L (48); E (8); G (4)	0.0568
23	E	7100913	99.94429	0	G (1886); V (1049); K (621); D (324); A (51); I (9); Q (3)	0.0557
26	K	7103367	99.97883	0	A (1013); T (227); R (197); E (31); N (18); Q (2); V (1)	0.0212
36	A	7102817	99.97109	0	N (1018); V (595); T (277); S (143); E (4); P (2)	0.0289
37	A	7103152	99.97581	0	T (1126); V (436); S (139); N,G,D (1)	0.0242
38	N	7102706	99.96953	0	A (1018); S (1003); T (85); D (34); H (9); V (1)	0.0305
42	K	7103043	99.97427	0	L (1018); R (407); E (349); I (15); T (14); N (8); H,Q (1)	0.0257
43	H	7102580	99.96775	0	K (1018); Y (983); Q (247); N (15); C (5); R (4); D (2); G,L (1)	0.0322
44	G	7100835	99.94319	0	V (1427); H (1018); E (902); R (667); A (7)	0.0568
45	G	7102503	99.96667	0	E (1134); R (1096); V (123)	0.0333
46	G	7103837	99.98545	0	C (367); S (330); D (276); V (44); R (2)	0.0146
47	V	7099222	99.92049	0	I (3249); A (1251); G (1025); F (108); L (1)	0.0795

48	A	7103654	99.98287	0	V (1073); T (71); S (47); G (7); E (4)	0.0171
49	G	7095928	99.87413	0	R (3624); V (2929); E (1327); A (1047); I (1)	0.1259
50	A	7096666	99.88452	0	V (3417); T (3210); G (1024); S (277); D (260); P,L (1)	0.1155
53	K	7102752	99.97018	0	N (1683); R (257); E (98); M (33); T (32); A (1)	0.0298
123	P	7103101	99.97509	0	A (1024); S (647); T (62); L (12); Q (8); K,R (1)	0.0249
124	L	7103660	99.98296	0	P (1024); V (157); I (9); S (4); Y,D (1)	0.0170
125	L	7104816	99.99923	0	V (31); I (9)	0.0008
126	S	7102832	99.9713	0	L (1514); A (239); P (227); T (29); K (12); Y (2); M (1)	0.0287
127	A	7098314	99.90771	0	V (3971); S (1395); T (1072); D (90); G (11); F (2); P (1)	0.0923
128	G	7101761	99.95623	0	D (1270); A (1037); S (648); C (83); V (55); L,R (1)	0.0438
129	I	7100006	99.93153	0	V (3473); G (1036); T (298); M (20); F,N (8); S (4); L (3)	0.0685
130	F	7103687	99.98334	0	I (1036); L (58); C (27); V (25); S (14); Y (9)	0.0167
131	G	7094562	99.8549	0	C (7758); F (1044); D (681); S (673); V (132); A (4); E,W (1)	0.1451
134	P	7085250	99.72384	0	L (9789); S (8052); D (1037); H (390); T (319); R (9); F (5); A (4); Y (1)	0.2762
152	A	7101292	99.94963	0	V (2026); L (1042); T (201); N (176); S (56);	0.0504

					D (32); G (30); P (1)	
153	V	7102750	99.97015	0	A (1305); I (519); F (264); G (11); D (6); L (1)	0.0299
154	F	7103650	99.98281	0	V (1050); L (84); S (50); C (13); I (8); Y (1)	0.0172
155	D	7101387	99.95096	0	E (1085); F (1042); G (831); Y (226); N (206); A (75); H,V (2)	0.0490
158	L	7092216	99.82188	0	F (10773); N (1042); I (528); R (146); P (91); H (24); S (19); V (12); C (5)	0.1781

Table 3.13. List of amino acid residues present at the binding site of SC-2 Mac1 domain of NSP3. The conserved residues are highlighted in green and intermediate conserved residues are highlighted in yellow (GISAID dataset). The AA position column shows the position of amino acid within the NSP3 Mac1 sequence, amino acid column shows the type of amino acid, number of sequences are the sequences having that exact amino acid, the conservation percent is the percentage of sequences having to amino acid as the reference sequence, gaps column shows if any gaps were inserted, mutations column shows the type of amino acid change along with the number of sequences having that mutated amino acid in brackets and mutational threshold column which shows the percentage of mutation.

NSP3 conservation across betacoronaviruses

The sequence conservation between the betacoronaviruses (SC-2, SC-1 and MERS-CoV) was high. Among the 34 amino acids within the binding site of Mac1 NSP3, 21 of them were 100% conserved and four amino acids were within the intermediate conservation threshold (Figure 3.29.) (Table 3.14.). The amino acids of the SC-2 Mac1 domain that were not found in SC-1 and MERS-CoV were the AA positions 19-23, 26, 154, 155 and 158.

```

SC_2_Mac1/1-167 1 VNSFSGYLKLTDNVYIKNADIVEEAKVKPTVVVNAANVYLKHGGGVAGALNKATNNA 58
SC_1/1-124 1 .....PMVIVNAANIHLKHGGGVAGALNKATNGA 29
MERS_CoV/1-124 1 .....ESVLVNAANTHLKHGGGIAAGINAAASKGA 29

SC_2_Mac1/1-167 59 MQVESDDYIATNGPLKVGGSCLVSGHNLAKHCLHVVGPVNVKGEDIQLLKSAYENFNQ 116
SC_1/1-124 30 MQKESDDYIKLNGPLTVGGSCLLSGHNLAKKCLHVVGPVNLNAGEDIQLLKAAYENFNS 87
MERS_CoV/1-124 30 VQKESDEYILAKGPLQVGDVLLQGHSLAKNILLHVVGPDARAKQDVSLLSKCYKAMNA 87

SC_2_Mac1/1-167 117 HEVLLAPLLSAGIFGADFIHSLRVCDTVRTNVYLAVF DKNLYDKLVSSFL 167
SC_1/1-124 88 QDILLAPLLSAGIFGAKPLQSLQVCVQTVRTQVYIAV..... 124
MERS_CoV/1-124 88 YPLVVTPLVLSAGIFGVKPAVSFDYLI REAKTRVLVVV..... 124

```

Figure 3.29. MSA of betacoronaviruses (SC-2, SC-1 and MERS-CoV) NSP3 Mac1 domain. Conserved amino acids are highlighted in purple whereas white colour represents amino acid variability.

Position	Amino acid	Number of sequences	Conservation percent	Mutations
36	A	3	100	
37	A	3	100	
38	N	3	100	
42	K	3	100	
43	H	3	100	
44	G	3	100	
45	G	3	100	
46	G	3	100	
47	V	2	66.6	I (1)
48	A	3	100	
49	G	3	100	
50	A	3	100	
53	K	2	66.6	A (1)
123	P	3	100	
124	L	3	100	
125	L	2	66.6	V (1)
126	S	3	100	
127	A	3	100	
128	G	3	100	
129	I	3	100	
130	F	3	100	
131	G	3	100	
134	P	3	100	
152	A	2	66.6	V (1)
153	V	3	100	

Table 3.14. List of amino acid residues present at the binding site of betacoronaviruses Mac1 domain of NSP3. The conserved residues are highlighted in green and intermediate conserved residues are highlighted in yellow. The AA position column shows the position

of amino acid within the Mac1 domain of NSP3 sequence, amino acid column shows the type of amino acid, number of sequences are the sequences having that exact amino acid, and the conservation percent is the percentage of sequences having to amino acid as the reference sequence (SC-2).

NSP3 Mac1 domain conservation across Group IV viruses

An MSA sequence alignment was carried out with group IV Mac1 domain sequences. There was not much sequence conservation between them (Figure 3.30.). But a few residues of binding site were 100% conserved, ALA36, ALA37, ASN38, GLY44, GLY46, LUE124, LEU125 and GLY128 (the residue numbers are as per SC-2 Mac1 NSP3). Since there was so little sequence conservation observed when performing an MSA of all these group IV viruses against the SC-2 Mac1 NSP3, we decided to compare these viruses as a part of PSA. The PSA was followed by a structural overlay of the group IV viruses and SC-2 Mac1 NSP3.

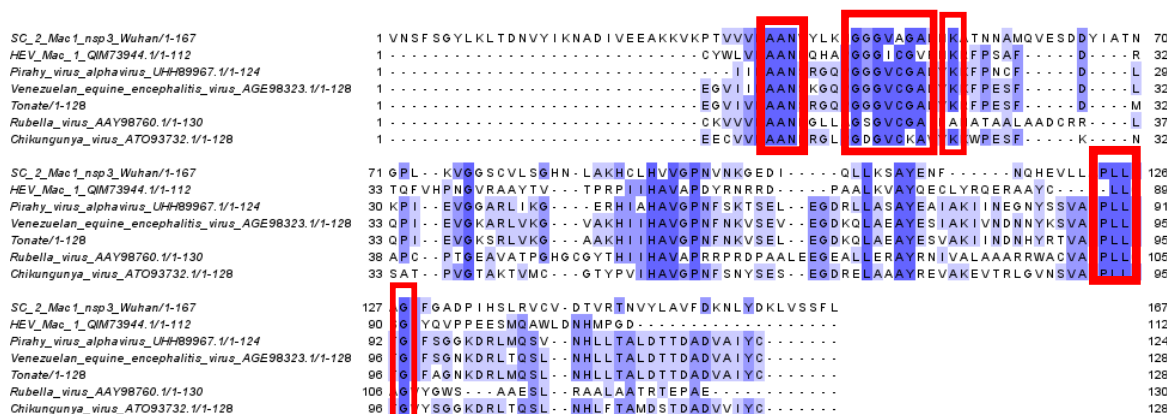


Figure 3.30. MSA of Mac1 domains of the group IV viruses (hepatitis E, Tonate virus, Venezuelan equine encephalitis virus, rubella virus, pirahy virus and chikungunya virus) with SC-2 Mac1 NSP3. The red boxes highlight the conserved amino acids within the RNA binding site. Conserved amino acids are highlighted in purple whereas white colour represents amino acid variability.

Group IV virus sequence and structural conservation to SC-2 NSP3

After comparing sequences and structures of a few group IV Mac1 domains against SC-2 Mac1 domain of NSP3, it was found that Chikungunya virus (CHIKV) NSP3 Mac1 domain has sequence/structural conservation with SC-2 Mac1 NSP3. Before moving to the comparison, sequence conservation was established for CHIKV nsP3 Mac1 sequences. A total of 280 CHIKV nsP3 sequences were used for generating the MSA. The mutational and

conservational thresholds selected were 0.71% and 0.0%, respectively. Since the dataset was small, a large amount of conserved amino acids were found. The top mutations in the dataset were THR77SER, GLY117ARG, SER48CYS, MET132LEU, ALA137VAL and SER153ALA (Figure 3.31.).

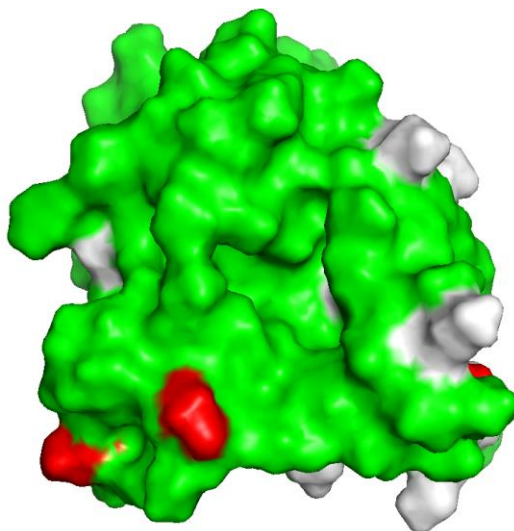


Figure 3.31. Conservational and mutational analysis of CHIKV nsP3 Mac1 (PDB: 7H8D) (a) Visualisation of conserved (green), mutated (red) and intermediate conserved (white) amino acid residues.

A PSA analysis was done between the SC-2 Mac1 domain of NSP3 and CHIKV nsP3 Mac1 domain. The amino acid identity between the two sequences was found to be 25% (36 amino acids) and the amino acid similarity was found to be 38.6% (54 amino acids) (Figure 3.32.). The identical amino acids are exactly the same amino acids that are present in the two sequences whereas similar amino acids are the ones that have similar chemical properties and exert similar functions. There were 14 amino acid positions that were identical and 11 amino acid positions that were similar within the binding sites of these proteins (Table 3.15.).

SC-2 AA position	SC-2 amino acid	Type of substitution	CHIKV amino acid
36	A (ALA)	identical AA	A (ALA)
37	A (ALA)	identical AA	A (ALA)
38	N (ASN)	identical AA	N (ASN)
42	K (LYS)	similar AA	L (LEU)
43	H (HIS)	similar AA	H (HIS)
44	G (GLY)	identical AA	G (GLY)
45	G (GLY)	similar AA	D (ASP)
46	G (GLY)	identical AA	G (GLY)
47	V (VAL)	identical AA	V (VAL)

48	A (ALA)	similar AA	C (CYS)
49	G (GLY)	similar AA	K (LYS)
50	A (ALA)	identical AA	A (ALA)
53	K (LYS)	identical AA	K (LYS)
123	P (PRO)	identical AA	P (PRO)
124	L (LEU)	identical AA	L (LEU)
125	L (LEU)	identical AA	L (LEU)
126	S (SER)	identical AA	S (SER)
127	A (ALA)	similar AA	T (THR)
128	G (GLY)	identical AA	G (GLY)
129	I (ILE)	similar AA	V (VAL)
130	F (PHE)	similar AA	Y (TYR)
131	G (GLY)	identical AA	G (GLY)
134	P (PRO)	similar AA	S (SER)
152	A (ALA)	similar AA	Y (TYR)
153	V (VAL)	similar AA	C (CYS)

Table 3.15. List of similar and identical amino acids present within the Mac1 binding site of SC-2 NSP3 and CHIKV nsP3. The AA position column shows the position of the amino acid within the Mac1 sequence (reference-SC2), the SC-2 amino acid column shows the type of amino acid present in the SC-2 sequence, the type of substitution column shows the similar/identical amino acids between the two sequences, and the CHIKV amino acid column shows the type of amino acid present in the CHIKV sequence.

```

# Length: 140
# Identity:      36/140 (25.7%)
# Similarity:   54/140 (38.6%)
# Gaps:         28/140 (20.0%)
# Score: 67.5
#
#
#=====
SC-2          1 PTVVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVG-GS      49
               ...|||||.....|.|||.:.|.....:      |.....| ..
Chikungunya   1 EECVVNAANPRGLPGDGVCKAVYKKWPESFK-----NSATPVGTAK      41

SC-2          50 CVLSGHNLAHKHCLHVVG---NVNKGEDIQLLKSAZENFNQHEVLLA---      93
               .|.:.| .....|.||| |...|.:.|.:.|.....|.
Chikungunya   42 TVMCG---TYPVIHAVGPNFSNYSESGDRELAAAYREVAKEVTRLGVNS      88

SC-2          94 ---PLLSAGIF--GAD----PIHSLRVCDTVRTNVYLAV      124
               |||||.:.: |.| .....|.:.|.:.|.....|.
Chikungunya   89 VAIPLLSTGVYSGGKDRLTQSLNHLFTAMDSTDADVVIYC      128

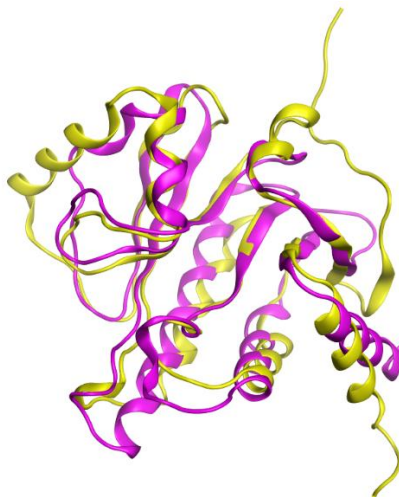
```

Figure 3.32. PSA between SC-2 Mac1 NSP3 and CHIKV Mac1 NSP3. (|) indicate identical amino acids, (:) indicate high similarity and (.) indicate weak similarity.

The structural overlay of SC-2 Mac1 NSP3 and CHIKV Mac1 NSP3 showed a good overall structural overlay with an RMSD of 2.2 Å (Figure 3.32. a). The structural and sequence overlay was then performed on the binding site of the Mac1 domain. The ADP-r binding

pocket of the two Mac1 domains were found to be highly conserved, with an overall RMSD of 1.3 Å (Figure 3.33. b). The binding site between the two proteins showed both sequence as well as structural conservation.

(a)



(b)

Pairwise RMSD Matrix:

Chains	1	2
1:7H8D.A		1.14
2:6WOJ.A	1.14	

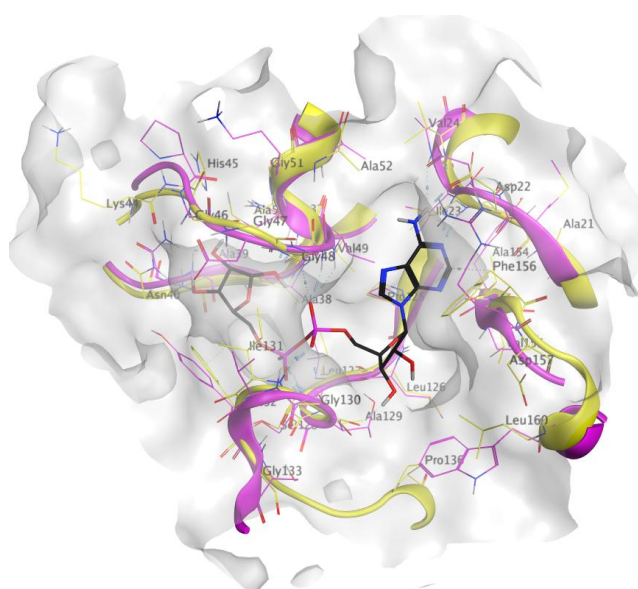
Overall RMSD: 1.139 Å

1: 7H8D.A VIRAL PROTEIN
vs.

2: 6WOJ.A VIRAL PROTEIN

RMSD = 1.139 Å

ALA 21	- MET 9	: 2.062	!
ASP 22	- ASP 10	: 1.780	
ILE 23	- ILE 11	: 0.700	
VAL 24	- ALA 12	: 1.333	
ASN 37	- ASN 21	: 0.280	
ALA 38	- ALA 22	: 0.401	
ALA 39	- ALA 23	: 0.838	
ASN 40	- ASN 24	: 1.257	
LYS 44	- LEU 28	: 0.940	
HIS 45	- PRO 29	: 1.950	
GLY 46	- GLY 30	: 1.167	
GLY 47	- ASP 31	: 1.523	
GLY 48	- GLY 32	: 1.080	
VAL 49	- VAL 33	: 0.515	
ALA 50	- CYS 34	: 0.742	
GLY 51	- LYS 35	: 0.484	
ALA 52	- ALA 36	: 1.006	
PRO 125	- PRO 107	: 0.389	
LEU 126	- LEU 108	: 0.330	
LEU 127	- LEU 109	: 0.590	
SER 128	- SER 110	: 0.772	
ALA 129	- THR 111	: 0.795	
GLY 130	- GLY 112	: 0.515	
ILE 131	- VAL 113	: 1.153	
PHE 132	- TYR 114	: 0.494	
GLY 133	- SER 115	: 0.875	
PRO 136	- LEU 121	: 1.888	
ALA 154	- TYR 142	: 0.761	
VAL 155	- CYS 143	: 0.748	
PHE 156	- ARG 144	: 1.691	
ASP 157	- ASP 145	: 1.534	
LEU 160	- TRP 148	: 1.830	



Targeting the most conserved residues that are present within the viral proteins such as helicases, proteases or polymerases has the risk of producing off-target interactions as these proteins are also present in humans that have a similar biological role. To check if there are any conserved residues between human PARP14 protein and SC-2 Mac1 NSP3, a pairwise sequence alignment was done. The human PARP14 protein was selected as it has an ADP-r binding site and is involved in ADP-ribosylation of target proteins. The PSA (Figure 3.34. a) suggested that there were 53 identical amino acids and 87 similar amino acids between the two proteins. The two proteins shared similar and identical amino acids at the ADP-r binding site. A structural overlay between human PARP14 and SC-2 Mac1 NSP3 also showed high structural similarity with an RMSD of 5.6 Å (Figure 3.34. b). Observing the high sequence and structural similarity between the two proteins, it is necessary to further investigate the off-target implications of the small molecule inhibitors targeting the Mac1 domain of SC-2 NSP3.

SC_2_Mac1	1	-----VNSFSGYLKLTDNVYIKNAD-----	20
		 : : : : : :	
PARP14	751	KQFFQDKARFYQSEIKRLFGCYIELQENEVMKEGGS	800
		AGKCFSRTVLAP	
SC_2_Mac1	21	----IVEA--KVKPTVVVNAANFYLKGGVGAGALNKA--NMQVESD	64
		:.. : : : : : : : : : : : : : : : : :	
PARP14	801	GVVLIVQGDGLRLPVVVVNASNEDLKYYGLAAALSKA	850
		APLQADCD	
SC_2_Mac1	65	DYIATNGPLKVGSGSVLSGHNLA-KHCLHVVGPNNVKGED---IQLLKSA	110
		 : : : : : : : : : : :	
PARP14	851	QIVKREGRLLPGNATISKAGLKPYHHVIHAVGPRWSGYEAPRCVYLLRRA	900
SC_2_Mac1	111	Y-----ENFNQHEVLLAPLLAGIFGADPI--SLRVCVDTVRT-----	147
		. : : : : : : : : : : : : : : : : :	
PARP14	901	VQLSLCLAKEYKYRSIAI-PAISGVFG----PLGRCVETIVSAIKENF	945
SC_2_Mac1	148	-----NVYLAVFKKLYDKL/SSFL-----	167
		: : : : : : : : : : : :	
PARP14	946	QFKKDGHCLEKIYLV--DVSEKTEAFAEAVKTVFKATLPDTAAPPGL	991

(b)

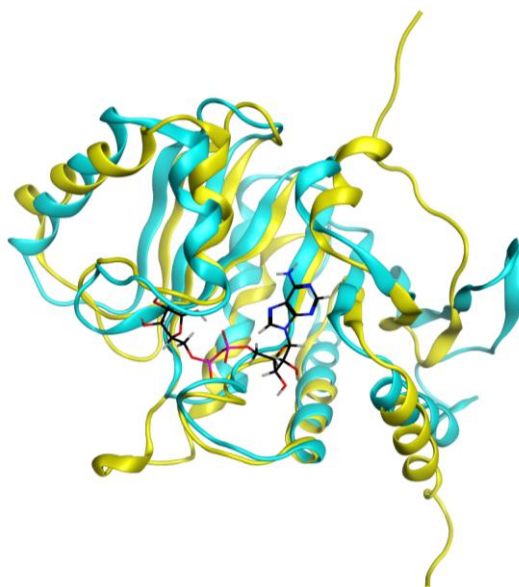


Figure 3.34. (a) PSA between SC-2 Mac1 NSP3 and human PARP14. The red boxes highlight the similar/identical amino acids within the ADP-r binding site (b) Structural overlay between human PARP14 (cyan) (PDB: 5QIA) and SC-2 Mac1 NSP3 (PDB: 6WOJ, ADP-r in black) (yellow).

(B) NSP3 cheminformatics analysis

Protein overlay and binding site analysis

The CACHE #Challenge3 was focussed on targeting the ADP-r binding pocket of Mac1 domain of SC-2 NSP3. The residues responsible for the ADP-r binding were highly conserved across beta-CoVs as mentioned in **the previous bioinformatics section** and was visualised on a 3D-protein X-ray crystal structure using MOE (Figure 3.35.).

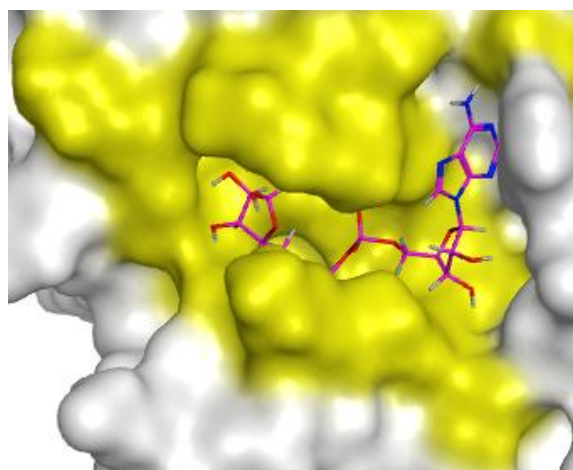


Figure 3.35. The highly conserved residues (yellow) present within the NSP3 Mac1 ADP-r binding pocket across the beta-CoVs. Pink colour: ADP-r (PDB ID: 7KQP).

There were many PDB submissions available with lead-like compounds co-crystalised within the binding pocket of Mac1 domain (Gahbauer et al., 2023). As a starting point, 22 PDB submissions with IC_{50} value of ≤ 10 μ M were chosen for protein structure and binding site analysis.

The 22 protein structures were aligned and overlaid using MOE to check how similar these structures were. The average RMSD of the overlay was 0.23 Å, suggesting that there was very little movement between these X-ray structures (Figure 3.36.). Residue positions 101-107, 128-136 and 156-159 deviated with an RMSD value 0.5-0.8 Å. There were a few AA residues present within the binding pocket that moved but with very small deviations.

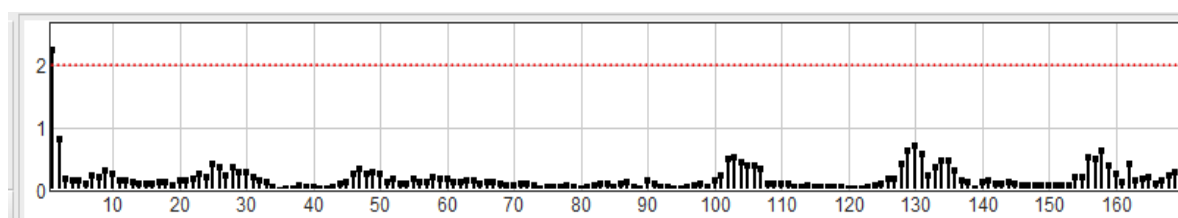
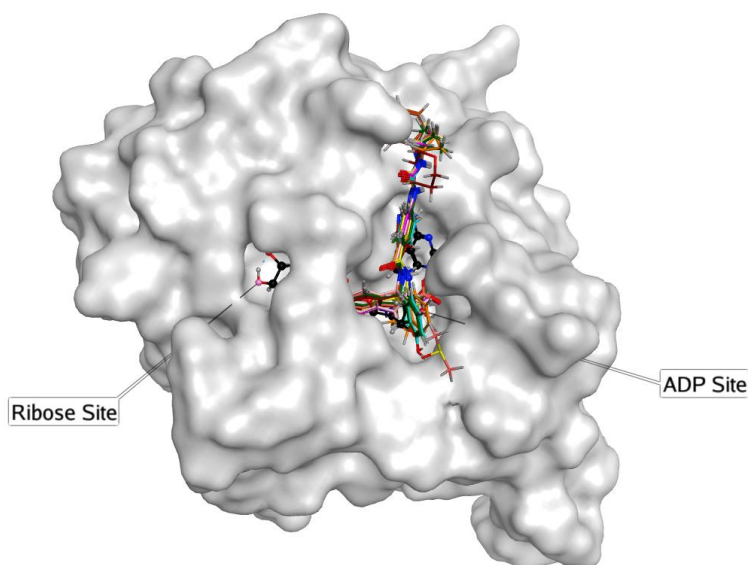
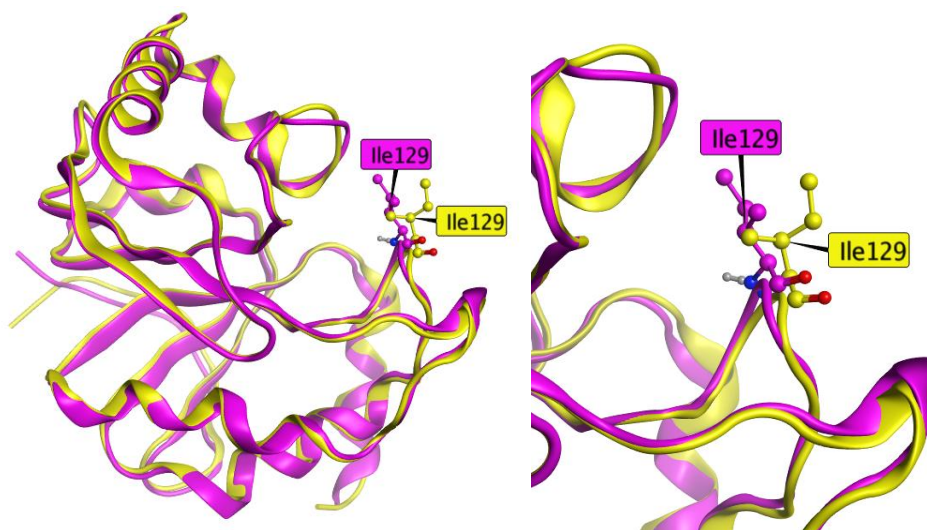


Figure 3.36. The structural overlay of 22 X-ray crystal structures with lead-like molecules in the Mac1 domain of SC-2 NSP3 along with RMSD plots (y-axis) of each amino acid residue position (x-axis). The average RMSD is plotted in red dotted line.

There were two different conformations of protein observed; (1) open tunnel (7KQP) and (2) closed tunnel (6WOJ) due to the presence of two conformations of isoleucine 129 (Figure 3.37.).

(a)



(b)

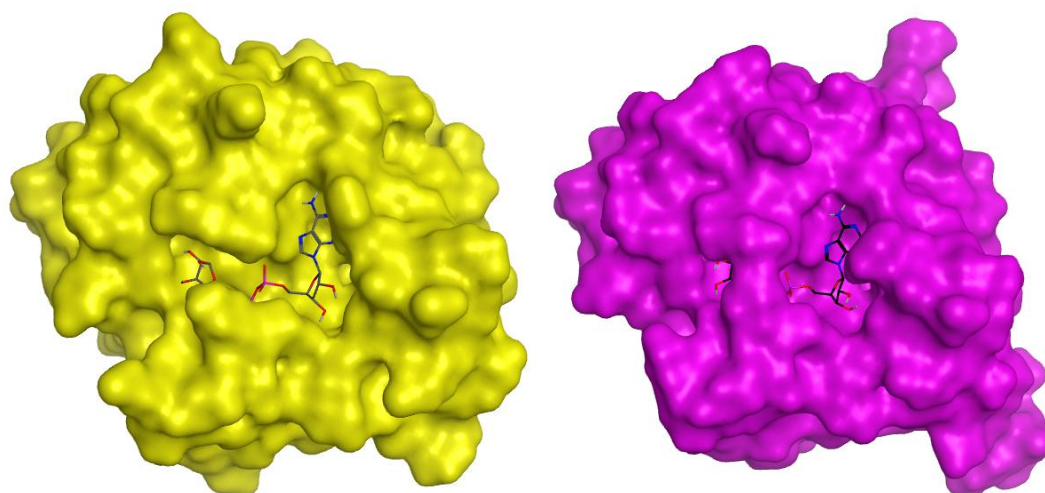


Figure 3.37. (a) Overlaid structure of 7KQP and 6WOJ with two conformations of isoleucine residue at 129 positions. (b) Open tunnel (yellow- 7KQP) and closed tunnel (pink-6WOJ) conformation with ADP-r co-crystallised within the binding pocket of NSP3 Mac1 domain.

The binding site was observed using the SiteView option on MOE. The AAs present within range of 4.5 Å of the binding pocket were noted. MOE SiteFinder was able to predict the

sites where the ADP-r and actives were co-crystalised (Table 3.16.). SiteFinder was also used for predicting different binding pockets other than the ADP-r and I binding sites.

Site	PDB ID's	Site Finder
ADP-r binding pocket	7KQP, 6WOJ	ALA19 ASP20 ILE21 GLY45 GLY46 VAL47 GLY49 ALA50 PRO123 LEU124 ALA127 ALA152 VAL153 PHE154 ASP155 LEU158 ASN35 ALA36 ALA37 ASN38 LYS42 HIS43 GLY44 GLY45 GLY46 VAL47 ALA48 GLY49 VAL93 VAL94 GLY95 LEU124 LEU125 SER126 ALA127 GLY128 ILE129 PHE130
22 actives site	5SP7, 5SPA, 5SPB, 5SPH, 5SQ5, 5SQ6, 5SQ9, 5SQC, 5SQH, 5SQJ, 5SQW, 5SQX, 5SQI, 5SRY, 5SRZ, 5SS9, 5SQD, 5SQY, 5SSM, 5SSP, 5SRL, 5SQO	ALA19 ASP20 ILE21 VAL22 ALA36 GLY45 GLY46 VAL47 ALA48 GLY49 ALA50 LYS53 ALA54 PRO123 LEU124 LEU125 SER126 ALA127 GLY128 ILE129 PHE130 GLY131 ALA132 PRO134 ALA152 VAL153 PHE154 ASP155 LEU158 ASN35 ALA36 ALA37 ASN38 LEU41 LYS42 HIS43 GLY44 GLY45 GLY46 ALA48 ASN97 LYS100 ILE129 PHE130

Table 3.16. The AA residues present at the two sites (ADP-r and actives) of SC-2 NSP3 along with the residues predicted by SiteFinder. The binding site was described to be within 4.5 Å. The site finder resulted in the amino acids and their positions for the two sites.

The dummy atoms were present within the binding pockets of the 22 lead-like molecules and ADP-r. Most of the AA residues present within these binding pockets were found to be conserved in SC-2 Mac1 NSP3 variants (GISAD dataset-green) as well as in betacoronaviruses (blue) (Figure 3.38.). The unhighlighted ones were within the intermediate conservation threshold.

(a)

ALA19 ASP20 ILE21 GLY45 GLY46 VAL47 GLY49 ALA50 PRO123 LEU124 ALA127 ALA152
 VAL153 PHE154 ASP155 LEU158 ASN35 ALA36 ALA37 ASN38 LYS42 HIS43 GLY44 GLY45
 GLY46 VAL47 ALA48 GLY49 VAL93 VAL94 GLY95 LEU124 LEU125 SER126 ALA127 GLY128
 ILE129 PHE130

ALA19 ASP20 ILE21 GLY45 GLY46 VAL47 GLY49 ALA50 PRO123 LEU124 ALA127 ALA152
 VAL153 PHE154 ASP155 LEU158 ASN35 ALA36 ALA37 ASN38 LYS42 HIS43 GLY44 GLY45
 GLY46 VAL47 ALA48 GLY49 VAL93 VAL94 GLY95 LEU124 LEU125 SER126 ALA127 GLY128
 ILE129 PHE130

(b)

ALA19 ASP20 ILE21 VAL22 ALA36 GLY45 GLY46 VAL47 ALA48 GLY49 ALA50 LYS53
 ALA54 PRO123 LEU124 LEU125 SER126 ALA127 GLY128 ILE129 PHE130 GLY131 ALA132 PRO134
 ALA152 VAL153 PHE154 ASP155 LEU158 ASN35 ALA36 ALA37 ASN38 LEU41 LYS42 HIS43
 GLY44 GLY45 GLY46 ALA48 ASN97 LYS100 ILE129 PHE130

ALA19 ASP20 ILE21 VAL22 ALA36 GLY45 GLY46 VAL47 ALA48 GLY49 ALA50 LYS53
 ALA54 PRO123 LEU124 LEU125 SER126 ALA127 GLY128 ILE129 PHE130 GLY131 ALA132 PRO134
 ALA152 VAL153 PHE154 ASP155 LEU158 ASN35 ALA36 ALA37 ASN38 LEU41 LYS42 HIS43
 GLY44 GLY45 GLY46 ALA48 ASN97 LYS100 ILE129 PHE130

Figure 3.38. (a) ADP-r binding site amino acid residues with highlighted conserved amino acids. (b) 22 lead-like molecules binding site amino acid residues with highlighted conserved amino acids. The conserved residues from the GISAID dataset are highlighted in green whereas the conserved residues from the betacoronaviruses dataset are highlighted in blue.

Virtual screening (VS) workflow

After the SiteFinder dummy atom placements, the entire binding pocket of Mac1 was subdivided into six smaller pockets (Table 3.17.). Each pocket was colour coded and sdf files were generated for fastROCS screening (Figure 3.39.).

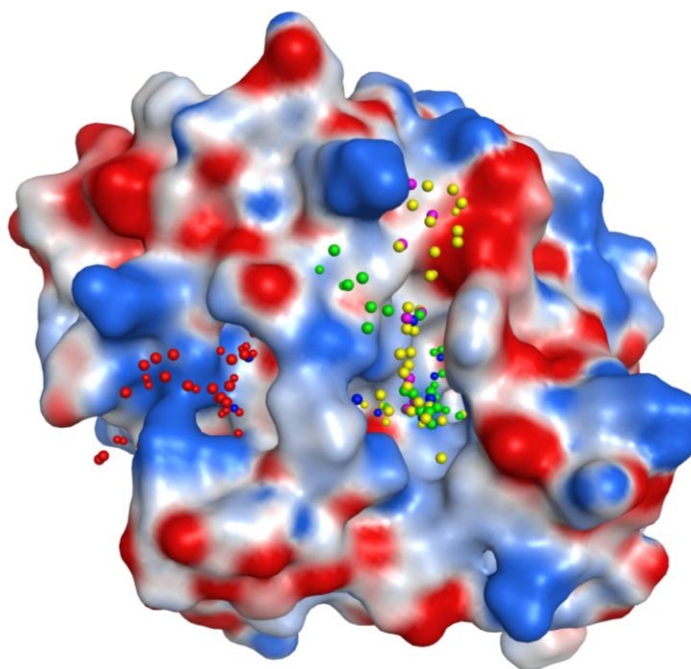


Figure 3.39. Placement of colour coded dummy atoms representing the six pockets on the Mac1 domain of NSP3. Protein electrostatic surface White: Hydrophobic AA, Red: Negative charge AA, Blue: Positive charge AA.

Pocket	Dummy atoms in 3D	Dummy atoms in 2D
Pocket 1 Site where actives are co-crystalised		
Pocket 2 Large (L) shaped pocket, which is an extension of pocket 1 covering half of the tunnel		

Pocket 3 Site adjacent to actives		
Pocket 4 Proximal ribose binding site		
Pocket 5 ADP-r binding pocket with closed tunnel conformation		
Pocket 6 ADP-r binding pocket with open tunnel conformation		

Table 3.17. Six different pockets with dummy atom placements on the Mac1 domain of NSP3. The dummy atoms are depicted in both 3D and 2D formats. In the 2D representation, the oxygen atoms are coloured red to indicate hydrophilic cavity points, while the carbon atoms are shown in black to represent hydrophobic cavity points.

1. The first round of screening was **fastROCS** volume query search on the Enamine database. Hits after the screening of each pocket mapped well with the dummy atom query (Figure 3.40.). The fastROCS screening resulted in total 450,000 hits for all the six pockets.

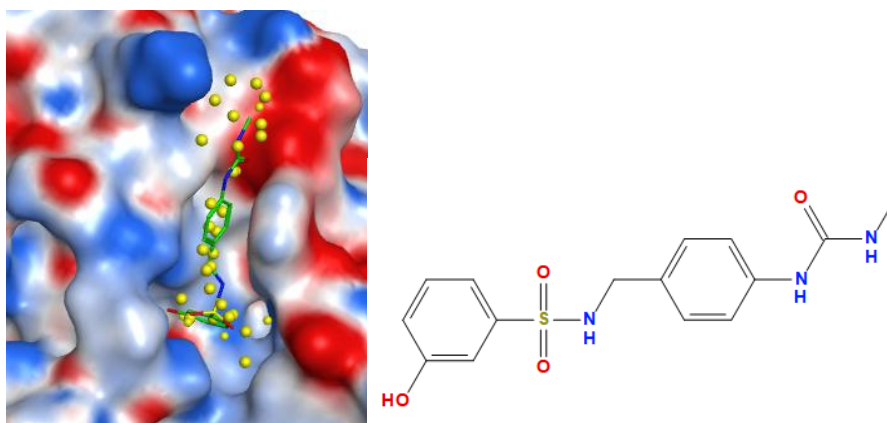
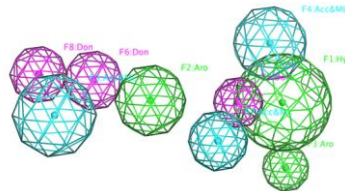
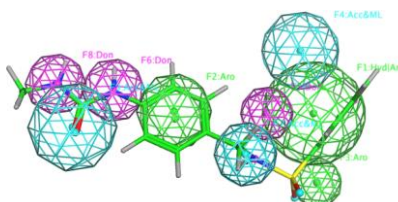
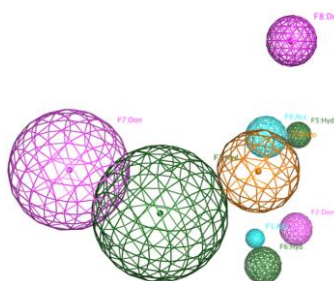
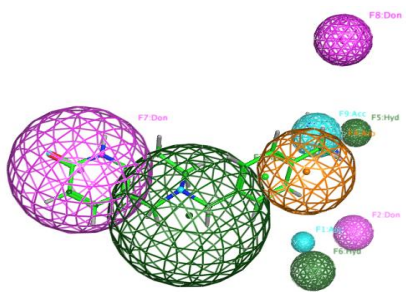
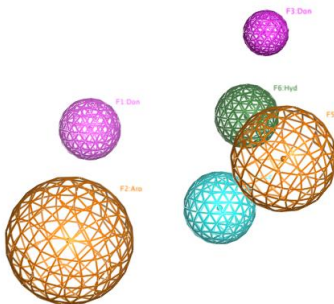
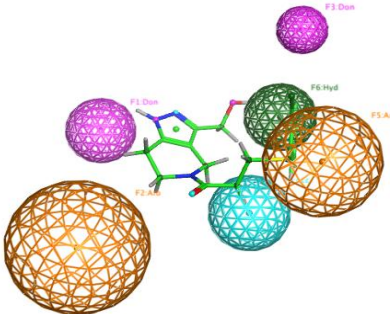
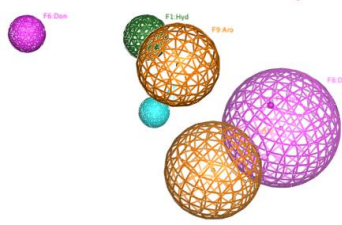
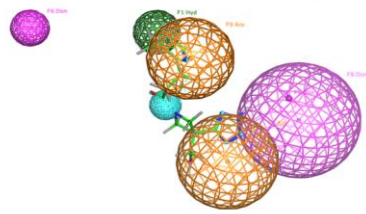


Figure 3.40. One of the hits PV-003713509148 (green) of pocket 1 mapping well with the dummy atoms (yellow) on the Mac1 domain of NSP3. Protein electrostatic surface White: Hydrophobic AA, Red: Negative charge AA, Blue: Positive charge AA.

2. The hits from the fastROCS screening were the input into the next round of VS i.e., **pharmacophore** search. The overlaid structure of 22 actives was used to generate a consensus pharmacophore with MOE. It resulted in a nine features pharmacophore query. A partial search was carried out with at least seven features and two essential acceptor features (F4 and F9). The consensus pharmacophore query was validated by screening through a haystack consisting of the 22 actives and 1100 decoys. The output resulted in 13 actives as hits thus confirming that the pharmacophore features cover the chemical description of the actives. MoPBS was used for rest of the pockets to generate the pharmacophore queries. For pocket 2 a nine features pharmacophore query with partial search was carried out with at least five features and two essential features i.e., one donor and one acceptor (F7 and F9). Pocket 3 had a six features pharmacophore query, and the search was carried out by ignoring one donor feature (F3) as it was outside the targeted site. Pocket 4 had a nine features pharmacophore query, and the search was carried out by ignoring three donor features (F3, F5 and F6). Pocket 5 had a seven features pharmacophore query, and the search was carried out by ignoring two features i.e., one aromatic and one donor feature (F3 and F4). Lastly, pocket 6 had a seven features pharmacophore query, and a partial search was carried out with at least five features.

This partial search is carried out when the actual query search does not result in any output. The essential and ignored features are assigned after carefully inspecting the binding site. The output of the pharmacophore searches for all the six pockets resulted in total 8,050 hits. The hits mapped well with the pharmacophore query (Table 3.18.).

The hit lists were filtered using the MOE database viewer which resulted in 7,476 hits after the removal of molecules having PAINS, carboxylic acid moieties, molecular weight > 400 g/mol, and SlogP > 3.5.

Pockets	Pharmacophore feature queries	Hit mapping with pharmacophore query
Pocket 1		
Pocket 2		
Pocket 3		
Pocket 4		

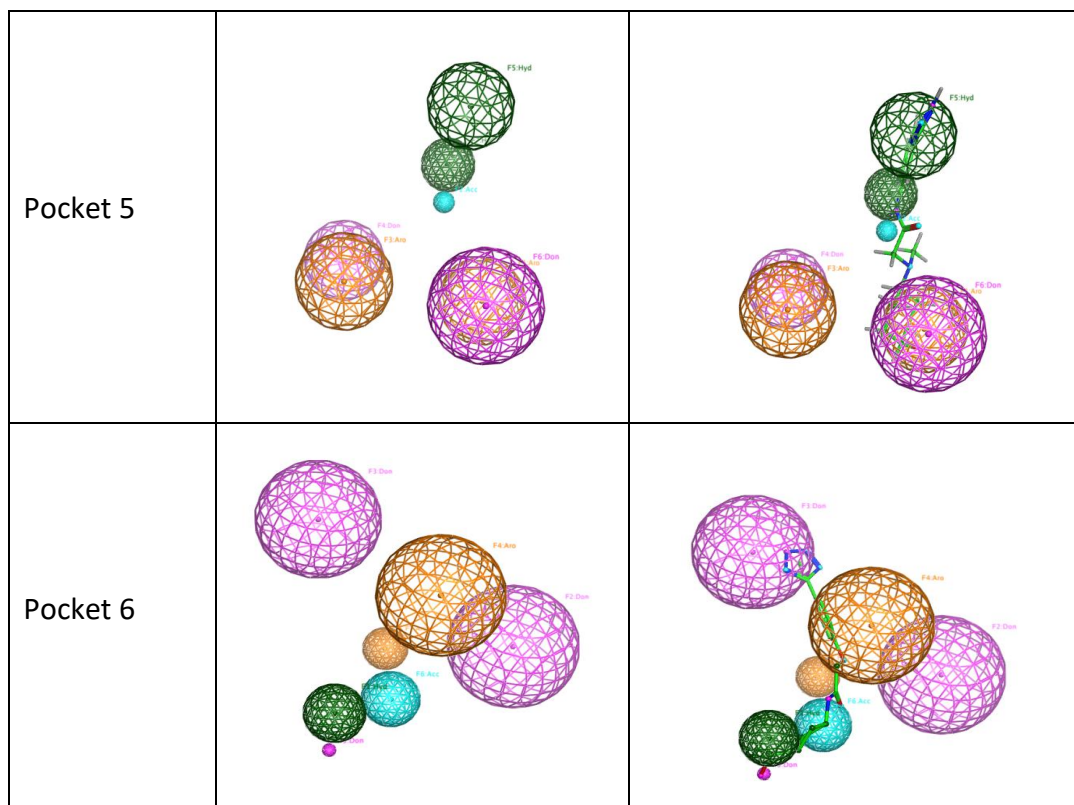


Table 3.18. Pharmacophore feature query for each pocket of the Mac1 domain of NSP3 and hits mapping with the features. Hyd (dark green)- Hydrophobic feature, Don (pink)- Hydrogen bond donor feature, Aro (orange)- Aromatic ring feature, Acc (cyan)- Hydrogen bond acceptor feature.

- The last round of VS was **molecular docking**. The hit lists after filtering were used for docking at six pockets. Before starting the docking VS, the efficiency of the FRED docking software was estimated using DataPype (Khan et al., 2023). The haystack consisting of 22 actives and 1100 decoys was screened and docked at pocket 1. It was found that FRED was able to identify and rank the actives at the top of the hit list with a good ROC-AUC of 0.88 and the graph was plotted using Pipeline Pilot (Figure 3.41.).



Figure 3.41. Shows the ROC-AUC curve with a score of 0.88. The x-axis shows rate of false positive, and y-axis shows rate of true positives.

The output of docking resulted in 7,296 hits for all the six pockets. The molecules from these hit lists were then clustered together (pocket 1-40 clusters, pocket 2- 40 clusters, pocket 3- 25 clusters, pocket 4- 25 clusters, pocket 5- 10 clusters and pocket 6- 10 clusters) to form a diverse set of 150 representative molecules using Pipeline Pilot. The FRED docking score was in the range of -12.24 to -5.86 of these 150 hits. All the 150 hits at the end of VS fulfilled both the fastROCS volume and pharmacophore features queries. The best interaction poses (Figure 3.42.) between the hits and protein were obtained by docking. The hits were observed to make interactions with AA residues that were conserved across the NCBI, GISAID and betacoronaviruses datasets.

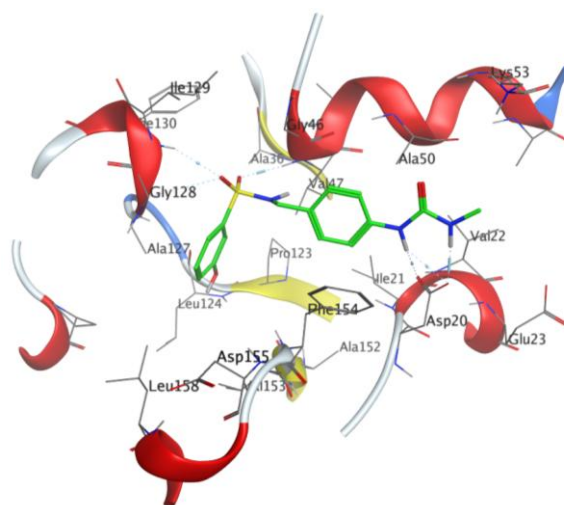


Figure 3.42. Docking of one of hits PV-003713509148 (green) at pocket 1 of Mac1 domain of NSP3 making interactions with highly conserved AA residues Ile129, Gly128, Val47.

After estimating the cost of purchase, 89 compounds were selected for synthesis by Enamine and subsequently sent to CACHE for experimental validation. Figure 3.43. summarises the entire tiered VS workflow of the MDG group.

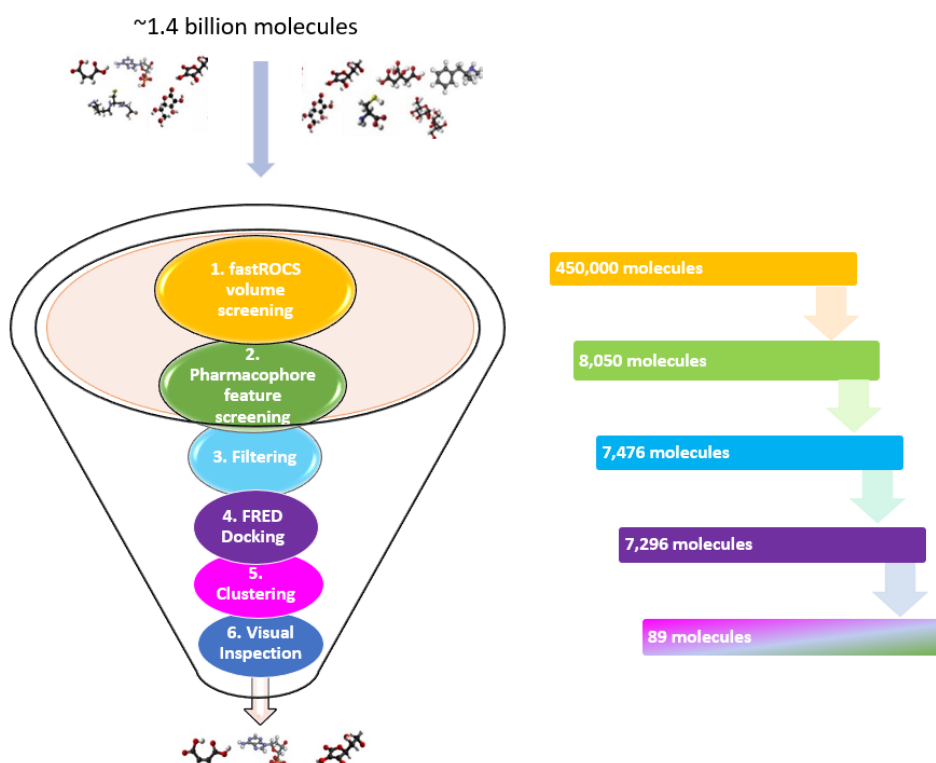


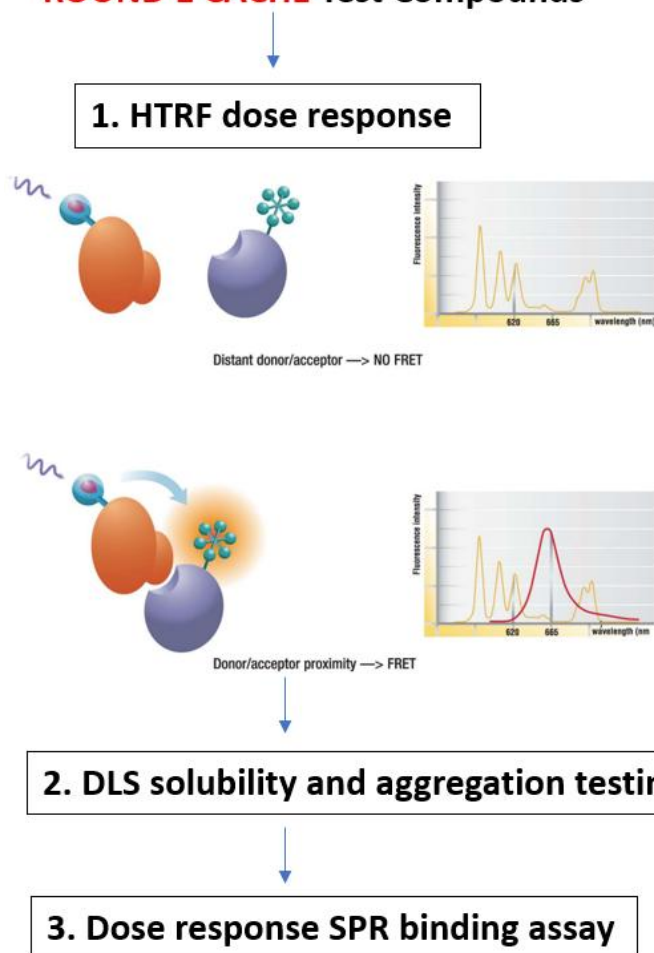
Figure 3.43. MDG tiered virtual screening workflow for SC-2 Mac1 NSP3. The workflow involved screening a database of approximately 1.4 billion molecules using fastROCS, followed by pharmacophore analysis, filtering, FRED docking, clustering, and ultimately visual inspection, which led to the identification of 89 molecules.

(C) Experimental validation results by the CACHE organisation

The CACHE organisation received submissions for predicted hits from 23 participants for challenge #3. The experimental validation was conducted in two rounds. For round 1, predicted hits from the submission (1,739 compounds) were screened in two independent runs at 100 μ M in a fluorescence-based Homogeneous Time Resolved Fluorescence (HTRF) competition assay (Gahbauer et al., 2023). The compounds were re-tested in the HTFR assay at 25, 50 and 100 μ M in duplicate in dose response experiment. In this round, the aggregation and solubility of the compounds were also measured by dynamic light

scattering (DLS). The compounds that showed clear dose dependent inhibition, >30% inhibition at 100 μ M, no fluorescence interference was advanced to subsequent SPR dose response experiment with concentrations 100 μ M, 50 μ M or 30 μ M. The SPR binding assay was validated with ADP-ribose and three positive controls (Z5010894404, Z5265428218 and Z5010894395) from Gahbauer et al., 2023. The compounds that had $K_d < 100 \mu$ M, > 30% of expected binding level, Chi2 < 10% Rmax (**CACHE cut-off**) and were advanced to second round of SPR assay but with 4% Dimethyl sulfoxide (DMSO) instead of 2% DMSO. The entire protocol is shown in Figure 3.44. The follow-up round 2 testing by CACHE involved the compounds to be tested again against target NSP3 and anti-target PARP14. The data for PARP14 will not be used to evaluate the compounds. The final compounds from round 1 that showed good binding to NSP3 Mac1 will also be solved by X-ray crystallography. As with CACHE #2, promising compounds were selected for a second round of hit expansions. As detailed below, none of our compounds progressed to this stage by CACHE, but we are bringing two of them through hit-to-lead studies.

ROUND 1 CACHE Test Compounds



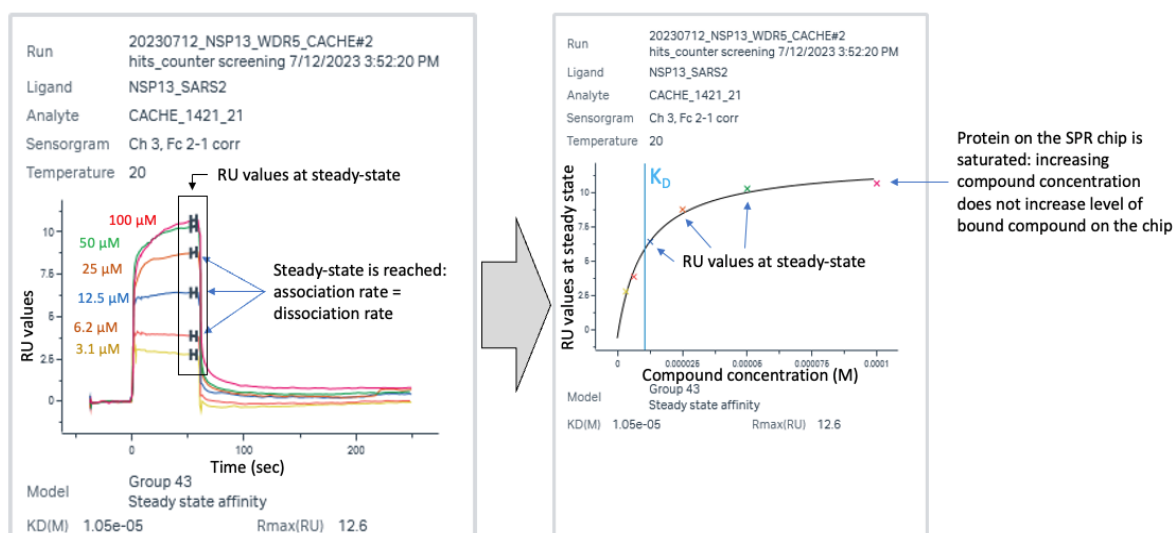


Figure 3.44. The experimental validation protocol involved various steps. (1) The HTRF competition assay was the first step, detecting ADP-r binding to the Mac1 domain and its displacement by the inhibitor. The His-tagged NSP3 Mac1 protein was labelled with a donor fluorophore, whereas the biotinylated ADP-r peptide was labelled with an acceptor fluorophore. If the proximity between them is less, there will be a strong HTRF signal, but if the inhibitor displaces the biotinylated ADP-r peptide, the HTRF signal will decrease. The Compounds that showed >30% inhibition/displacement of ADP-r were progressed to the next step. The schematic is taken from (Degorce, 2009). (2) The second step was determining DLS solubility and aggregation of the test molecules. If the aggregates are larger, the intensity ((kcnt/s) of the scattered light will increase and vice versa. In solubility determination, lower percentage laser power (%) indicates increased scattering from high particulate matter, resulting in poor solubility, and vice versa. (3) The third step involved SPR dose response binding assay. The compounds predicted in CACHE challenge are relatively weak binders, with rapid association (K_a) and dissociation (K_d) and do not contain sufficient kinetic information for reliable binding evaluation by kinetic fitting. Instead, the steady-state method is applied, where K_d is determined by plotting RU values measured at steady-state versus compound concentration. The SPR sensorgrams example graph (left) show response unit (RU) changes over time for increasing concentrations of the compound injected over immobilised SC-2 NSP3 Mac1 protein. The steady-state region, where association and dissociation rates are balanced, was used to determine binding levels at each concentration (different colours). The x-axis shows time (seconds) and the y-axis shows RU values. The steady-state RU values were

plotted against compound concentration (right) to generate a binding curve, from which the equilibrium dissociation constant (K_d) was calculated. Saturation of the response at higher concentrations indicates that the binding sites on the protein are fully occupied. The x-axis shows compound concentration (M) and the y-axis shows RU values at steady-state. The RU represents binding signal and R_{max} represents maximum RU (extrapolated by the software if steady state was not reached). The example SPR sensorgram was provided by the CACHE organisation.

In round 1, we submitted 89 compounds for experimental testing. The compounds were screened in a primary HTRF assay at 100 μ M against SC-2 NSP3. Five compounds showed >30% inhibition in the primary HTRF assay but only three compound passed the dose response HTRF assay (Table 3.19.). The three compounds were then subjected to solubility and aggregation testing by DLS and it was found that compounds CACHE3HI_1689_35 and CACHE3HI_1689_80 had high aggregation and poor solubility (Table 3.20.).

Compound	Inhibition %	HTRF-dose response
CACHE3HI_1689_13	32.6	fluorescence interference
CACHE3HI_1689_15	33.2	Passed HTRF
CACHE3HI_1689_20	81.4	fluorescence interference and no dose dependent response
CACHE3HI_1689_35	31.7	Passed HTRF
CACHE3HI_1689_80	57.6	Passed HTRF

Table 3.19. Compounds that showed >30% inhibition in the primary HTRF assay. These results were produced by the CACHE organisation.

Compound	% Laser power (for solubility)		Intensity (kcnt/s) (for aggregation)	
	100 μ M	50 μ M	100 μ M	50 μ M
CACHE3HI_1689_15	100	100	2622	1201
CACHE3HI_1689_35	33	65	4198	5730
CACHE3HI_1689_80	25	68	8503	5403

Table 3.20. DLS solubility and aggregation results of the compounds. The intensity is in (kcnt/s) and laser power is in (%) for both the concentrations of the test compounds. These results were produced by the CACHE organisation.

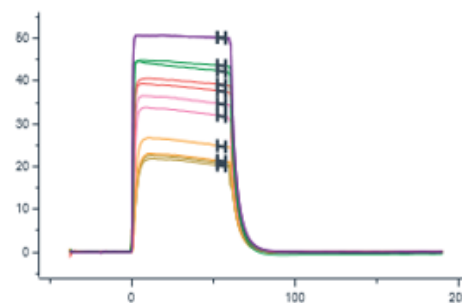
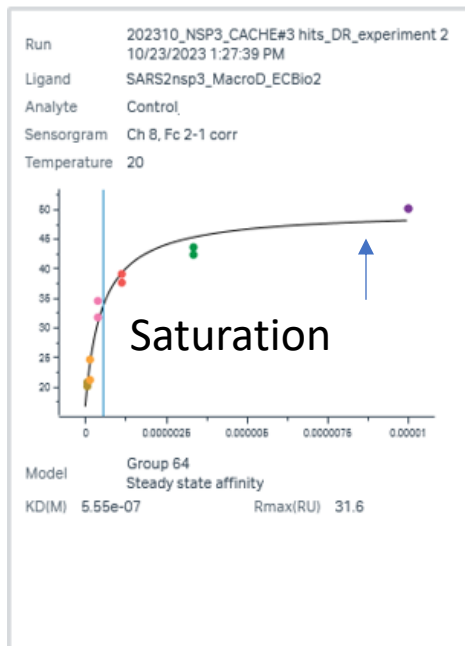
These three compounds were then subjected to SPR dose response experiments with 2% DMSO. The compound CACHE3HI_1689_80 was not selected for next round of testing as it missed the K_d cut-off value. Only CACHE3HI_1689_15 and CACHE3HI_1689_35 proceeded to another round of SPR dose response experiments with 4% DMSO. The CACHE also conducted the SPR dose response on a positive control compound (Figure 3.44. a) The compound CACHE3HI_1689_15 had K_d values of 328 μM and >1000 μM at 2% and 4% DMSO concentrations respectively. This compound was not selected for round 2 evaluation because it missed the K_d cut-off value. The compound CACHE3HI_1689_35 had K_d values of 4.84 μM and 2.6 μM at 2% and 4% DMSO concentrations respectively. This compound was also not selected for round 2 evaluation because it missed the % binding cut-off value, had a noisy sensorgram and DLS solubility testing indicated solubility issue (Table 3.21.) (Figure 3.45.).

Compound	2% DMSO			4% DMSO		
	K_d (μM)	% binding	Decision	K_d (μM)	% binding	Decision
CACHE3HI_1689_15	328	185.4	missed K_d cutoff ($K_d > 100$ μM)	>1000	>1000	missed K_d cutoff ($K_d > 100$ μM)
CACHE3HI_1689_35	4.84	19.7	% binding too low	2.6	1.2	% binding too low $> (>30\%)$, noisy sensorgram and DLS indicating solubility issue
CACHE3HI_1689_80	2200000	671501.3	missed K_d cutoff ($K_d > 100$ μM)	-	-	-

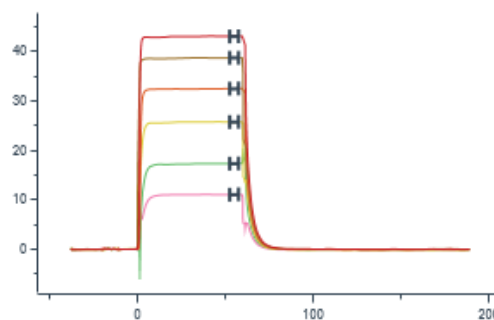
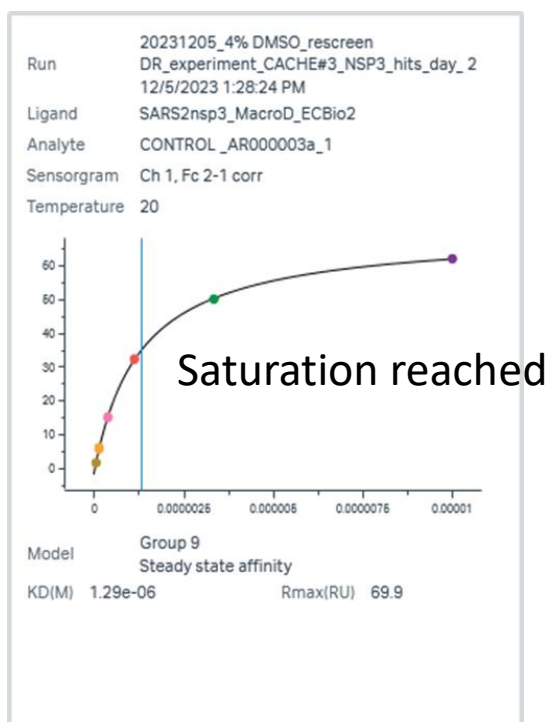
Table 3.21. SPR results of the compounds with 2% and 4% DMSO concentrations. For both the runs, K_d (μM) and % binding values were achieved, and a decision for further testing was made. None of the compounds were progressed to the next round of experimental validation. These results were produced by the CACHE organisation.

(a)

Screen at 2% DMSO containing buffer

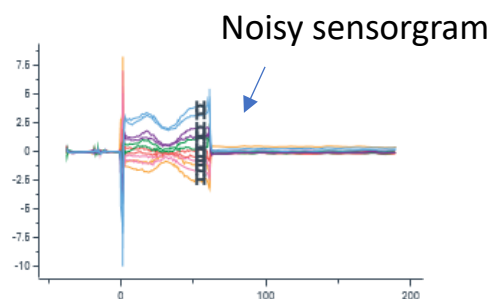
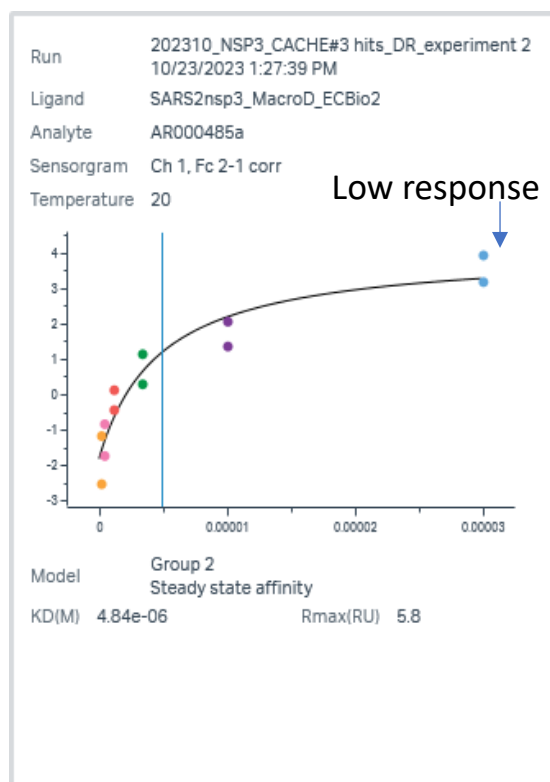


Screen at 4% DMSO containing buffer

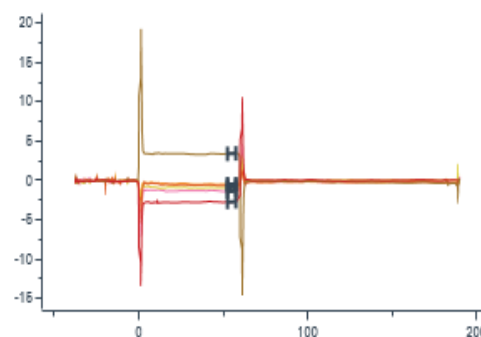
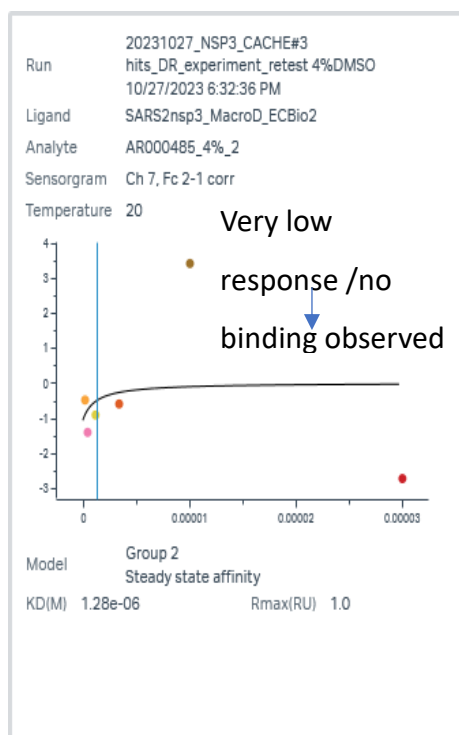


(b)

Screen at 2% DMSO containing buffer

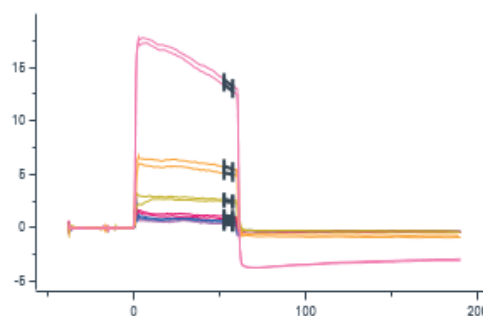
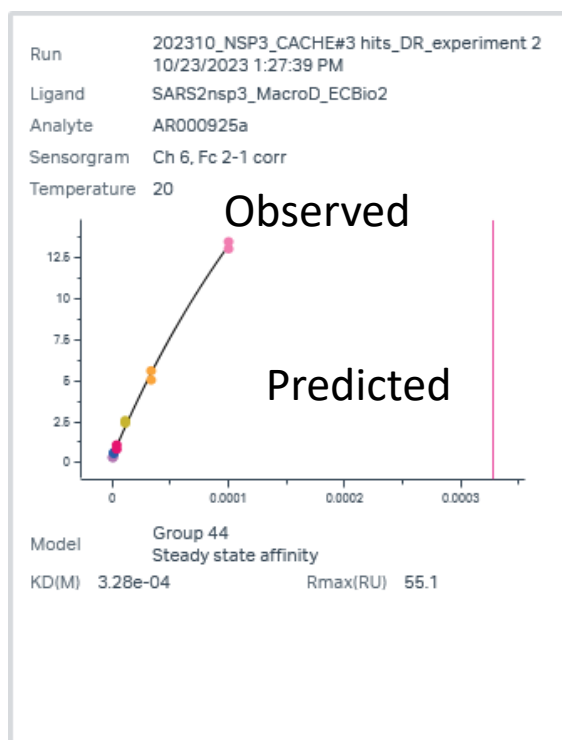


Re-screen at 4% DMSO containing buffer



(c)

Screen at 2% DMSO containing buffer



Re-screen at 4% DMSO containing buffer

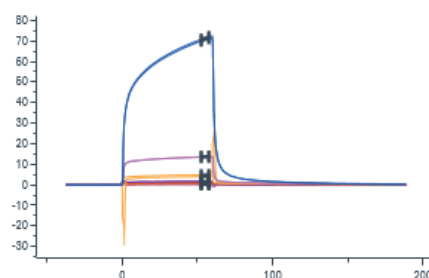
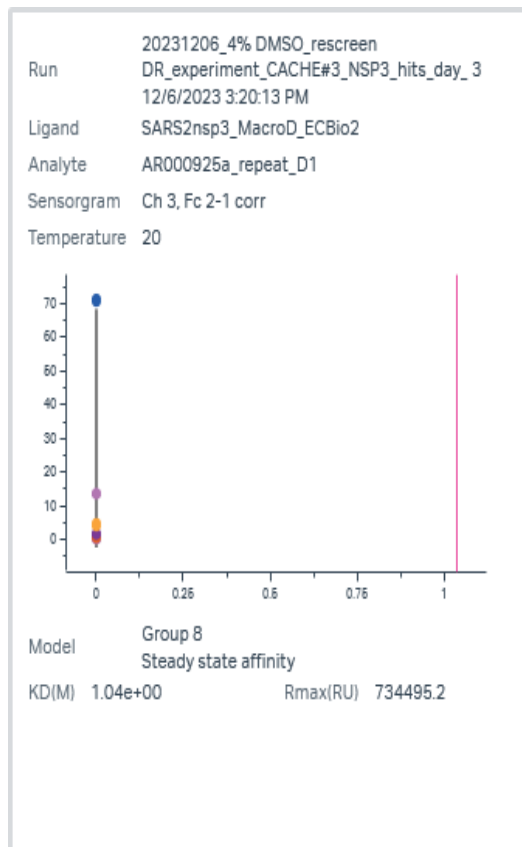
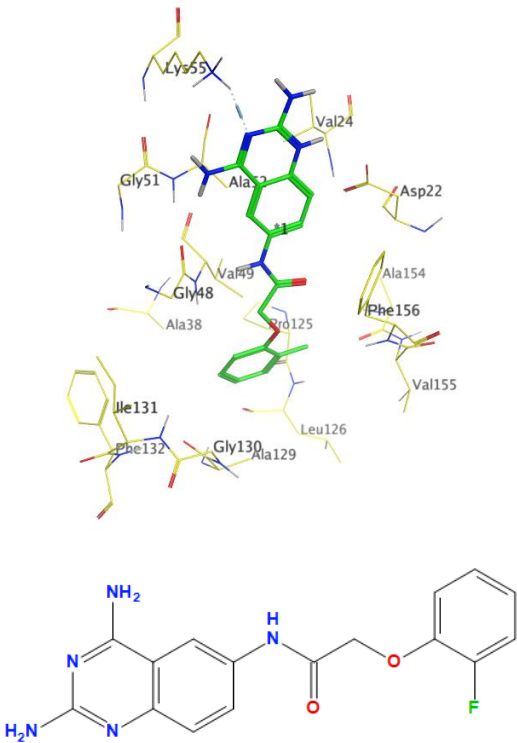
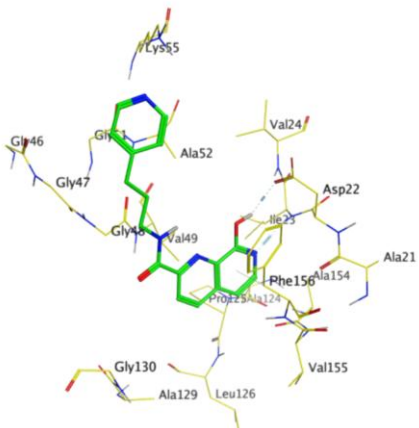


Figure 3.45. (a) The SPR sensorgrams show the response unit (RU) for the positive control compound at 2% and 4% DMSO. In both cases, the SPR sensorgrams (right) show the response unit (RU) on the y-axis and time (seconds) on the x-axis. Different concentrations (shown in various colours) of the compounds were injected over immobilised NSP3 Mac1 protein. The plateau of each curve indicates the steady-state binding level at that concentration. The graph on the left side shows corresponding steady-state response values plotted against compound concentration (x-axis) with RU at steady state (y-axis), which represents the amount of bound compound. The equilibrium dissociation constant (K_d , binding affinity) was calculated from the fitting binding curve, which shows saturation at high concentrations. (b) The SPR sensorgrams show the response unit (RU) for CACHE3HI_1689_35 compound at 2% and 4% DMSO. In both cases, the SPR sensorgrams (right) show the response unit (RU) on the y-axis and time (seconds) on the x-axis. Different concentrations of the compounds were injected over immobilised NSP3 Mac1 protein. The sensorgrams achieved were noisy for both the cases. The graph on the left side shows corresponding steady-state response values plotted against compound concentration (x-axis) with RU at steady state (y-axis), which represents the amount of bound compound. For both the cases a low response was observed. The equilibrium dissociation constant (K_d , binding affinity) was calculated from the fitting binding curve, which shows low response/no binding at high concentrations. (c) The SPR sensorgrams show the response unit (RU) for CACHE3HI_1689_15 compound at 2% and 4% DMSO. In both cases, the SPR sensorgrams (right) show the response unit (RU) on the y-axis and time (seconds) on the x-axis. Different concentrations of the compounds were injected over immobilised NSP3 Mac1 protein. The graph on the left side shows corresponding steady-state response values plotted against compound concentration (x-axis) with RU at steady state (y-axis), which represents the amount of bound compound. For both the cases extremely, weak binding was observed. The equilibrium dissociation constant (K_d , binding affinity) was calculated from the fitting binding curve, which shows very low-affinity or essentially non-specific binding at high concentrations. In these graphs, RU represents binding signal and R_{max} represents maximum RU (extrapolated by the software if steady state was not reached). These results were produced by the CACHE organisation.

The 2 hits from the round 1 experimental validation results are shown in Table 3.22. with the docked score and the site where they were docked. The two hits were subjected to molecular dynamics simulations to observe the stability of the interactions between them and binding site amino acid residues, which is discussed briefly in Chapter IV. Since the two hits are promising, initially 20 structural analogues of CACHE3HI_1689_35 were selected for synthesis by Enamine and are currently undergoing biological evaluation by a collaborator (Prof Choon Kim) with the aid of a Notre Dame-DCU Seed Grant in Precision Biomedical Technologies funding.

Compound	Docked pose	Docking score	Pocket
CACHE3HI_1689_15		-6.87	Pocket 1
CACHE3HI_1689_35		-8.7	Pocket 3

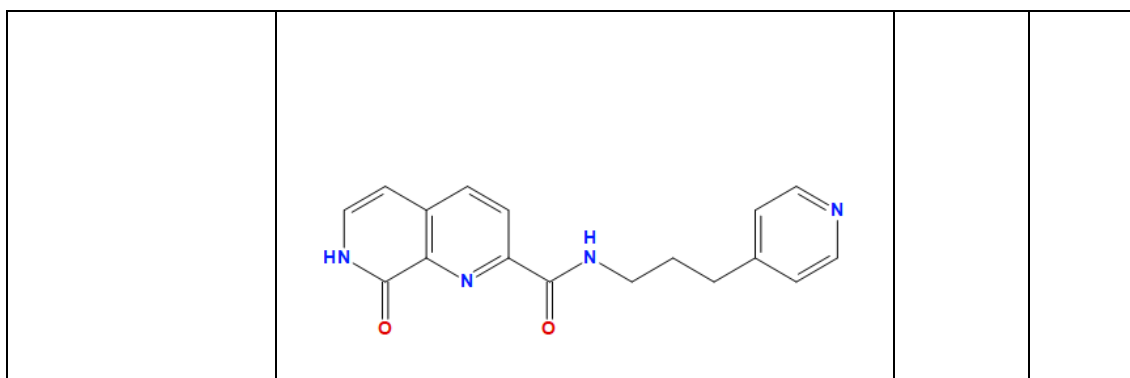


Table 3.22. Two hits (green) from round 1 experimental validation results with docked poses, docking score and site of docking (yellow). The dotted lines represent hydrogen bonds as well as hydrophobic interaction.

Protocol validation

The CACHE organisation released the list of compounds (actives) that showed binding to Mac1 NSP3 in round 1 and round 2 of experimental validation. To test and improve our VS protocol, we performed the entire workflow on these actives, as well as the compounds that did not show binding to Mac1 NSP3 as inactives. For both the rounds together, there were a total of 94 actives and 1787 inactives.

When we used our VS workflow on these datasets, all the actives were found as hits in fastROCS VS. However, the pharmacophore VS only identified three actives as hits for pocket 3 and two hits for pocket 6. For the rest of the pockets, no actives were identified as hits. Our pharmacophore VS stage was too stringent at filtering potential actives. The pocket which resembled the site where the 22 lead-like molecules were co-crystallised was pocket 1. The pharmacophore feature query was changed from a partial search with at least 7 features to at least 4 features for pocket 1. This adjustment of the pharmacophore features helped in identifying 73 actives. Lastly, the make receptor docking grid box size was increased for pocket 1 to cover the space of pocket 2 and pocket 3. After running the entire protocol again for this enlarged pocket 1, the active molecules were identified first and ranked at the top with the best docked scores (ROC-AUC 0.9) (Figure 3.46.). The validation protocol thus helped modify the VS workflow. This updated workflow was used for hit to lead progression of the two hits identified in round 1 of experimental validation.

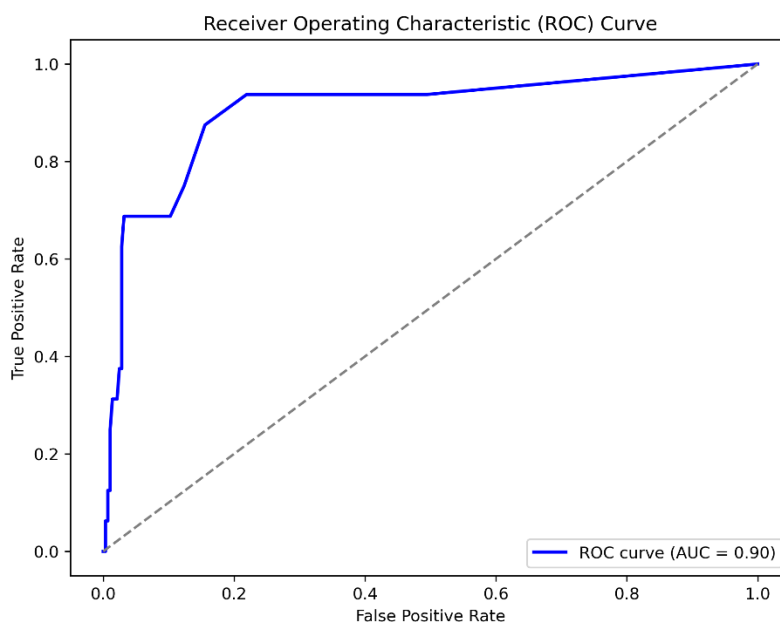


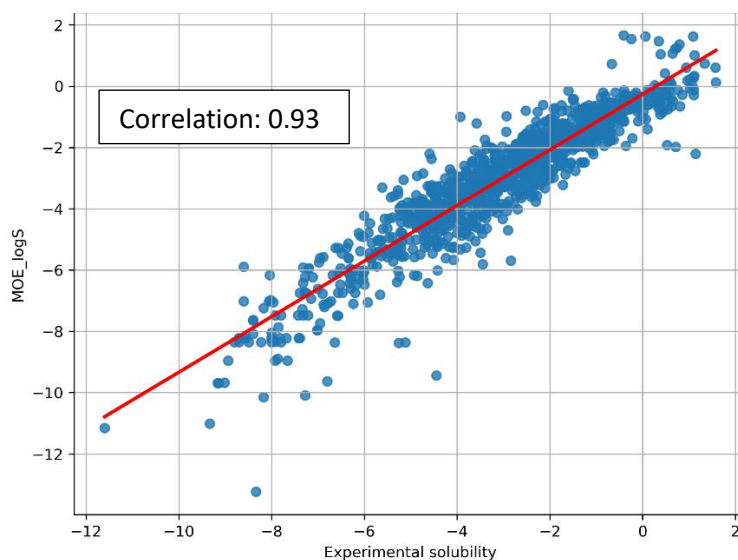
Figure 3.46. ROC-AUC curve of actives and inactives of Mac1 NSP3 from the CACHE experimental validation results. The x-axis shows false positive rate, and the y-axis show true positive rate.

Hit to lead progression

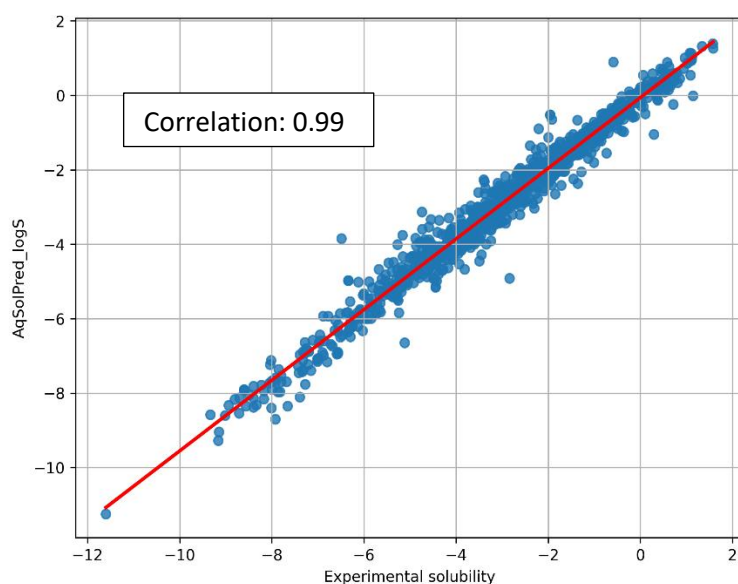
From the CACHE round 1 results, two hits, CACHE3HI_1689_15 and CACHE3HI_1689_35, were used for substructure searching using RDKit to find structural analogues from the Enamine REAL database. Multiple structural variations, using SMART notation, (**Appendix 6.3.**) were used for the substructure searching by making changes in the core structures of the two hits. The substructure searches of CACHE3HI_1689_15 and CACHE3HI_1689_35 resulted in 23,746 and 25,605 hits, respectively. The hit lists were then used for VS through the optimised pharmacophore query search and FRED molecular docking. The hits were screened at pocket 1; the pharmacophore search resulted in 14,295 hits for CACHE3HI_1689_15 and 19,851 hits for CACHE3HI_1689_35. The hits were then ranked as per the FRED docking score. In the experimental validation stage done by CACHE, it was found that CACHE3HI_1689_35 had solubility issues. To confirm that the structural analogues had good solubility, an initial benchmark study was carried out between MOE and AqSolPred to predict logS values of molecules having known experimental LogS values. The AqSolPred is a solubility prediction model which consists of 3 ML algorithms: neural nets, random forest, and XGBoost (Sorkun et al., 2020). The performance of MOE and AqSolPred was validated based on how well they predict logS when compared to the

experimental logS values. Two datasets were used; one had 1,128 molecules, and the other dataset had 3,942 molecules. To compare MOE and AqSolPred performances, correlation plots were generated. Both the datasets, MOE and AqSolPred showed good correlations with experimental data, but AqSolPred performed slightly better in predicting the solubility of the molecules (Figure 3.47.).

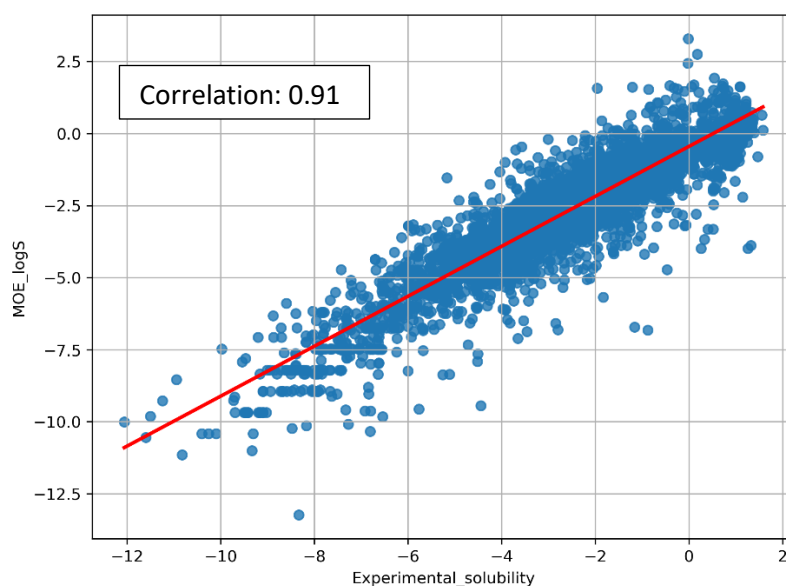
(a)



(b)



(c)



(d)

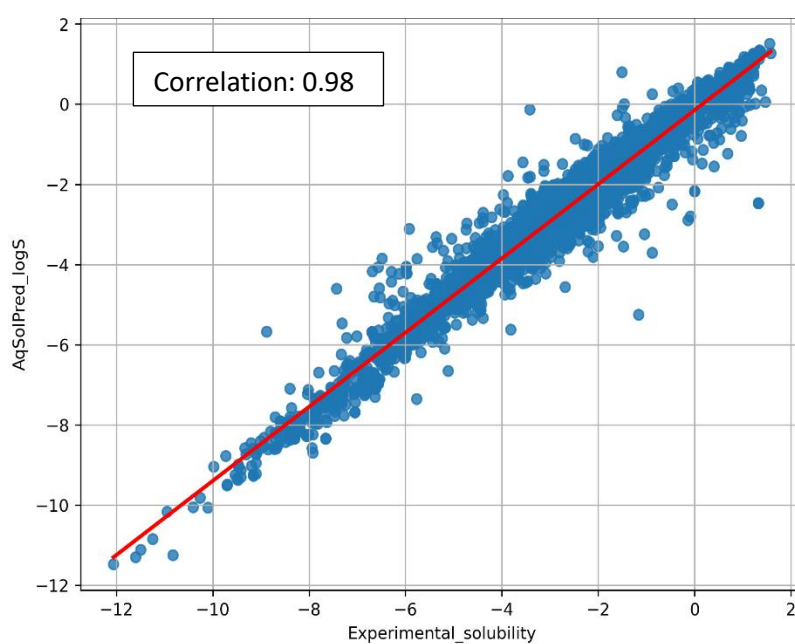
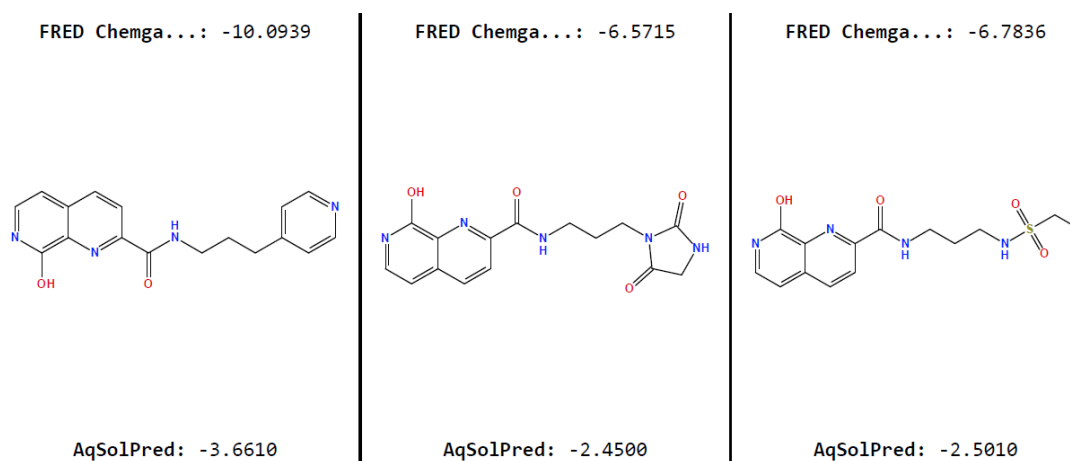


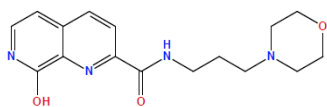
Figure 3.47. (a) Co-relation plot between the experimental logS values and MOE predicted logS values for 1,128 molecules dataset. The x-axis shows experimental solubility (logS) and the y-axis show predicted logS values by MOE. (b) Co-relation plot between the experimental logS values and AqSolPred predicted logS values for 1,128 molecules dataset. The x-axis shows experimental solubility (logS) and the y-axis show predicted logS values by AqSolPred. (c) Co-relation plot between the experimental logS

values and MOE predicted logS values for 3,942 molecules dataset. The x-axis shows experimental solubility (logS) and the y-axis show predicted logS values by MOE. (d) Correlation plot between the experimental logS values and AqSolPred predicted logS values for 3,942 molecules dataset. The x-axis shows experimental solubility (logS) and the y-axis show predicted logS values by AqSolPred.

The AqSolPred predicted LogS values of the CACHE3HI_1689_15 and CACHE3HI_1689_35 were -3.7 and -3.6, respectively. As per the AqSolPred model, compounds with 0 and higher solubility values are highly soluble; those in the range of 0 to -2 are soluble; those in the range of -2 to -4 are slightly soluble; and compounds less than -4 are insoluble. The two hits were in the slightly soluble range. After the FRED docking of the structural analogues, AqSolPred was used to predict the logS values of the molecules. The final hit list was ranked from best to worst solubility logS scores and was used to guide the selection of analogues for synthesis. For the first round of experimental testing (binding assay), 19 analogues were selected and purchased from Enamine along with the parent CACHE3HI_1689_35 molecule as a part of batch one (Figure 3.48.). The range of log S scores for the hits was -2.4 to -4.1. Three hits had logS scores in the range of -4.0 to -4.15; these were considered for testing as they made a high number of hydrogen bonds with the protein backbone. Another batch of 20 analogues of CACHE3HI_1689_15 will be tested after the in vitro data of the first batch is available and used to update our predictive models.

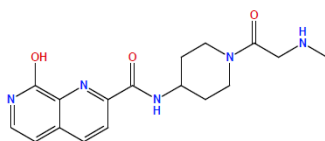


FRED Chemga...: -6.5276



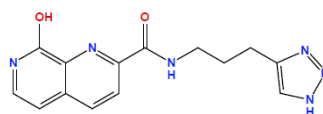
AqSolPred: -2.6620

FRED Chemga...: -7.5060



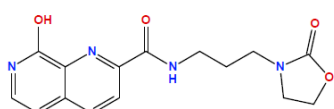
AqSolPred: -2.6880

FRED Chemga...: -8.1080



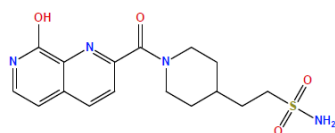
AqSolPred: -2.7830

FRED Chemga...: -7.3709



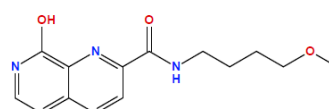
AqSolPred: -2.8080

FRED Chemga...: -7.4482



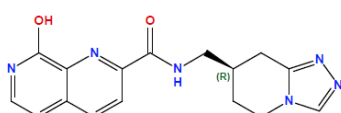
AqSolPred: -2.8220

FRED Chemga...: -7.1800



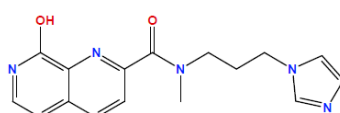
AqSolPred: -3.0140

FRED Chemga...: -6.4835



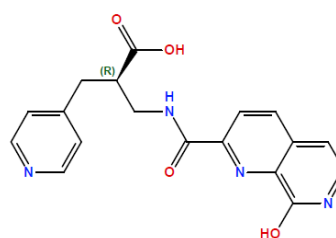
AqSolPred: -3.0270

FRED Chemga...: -7.8009



AqSolPred: -3.0620

FRED Chemga...: -9.7781



AqSolPred: -3.5040

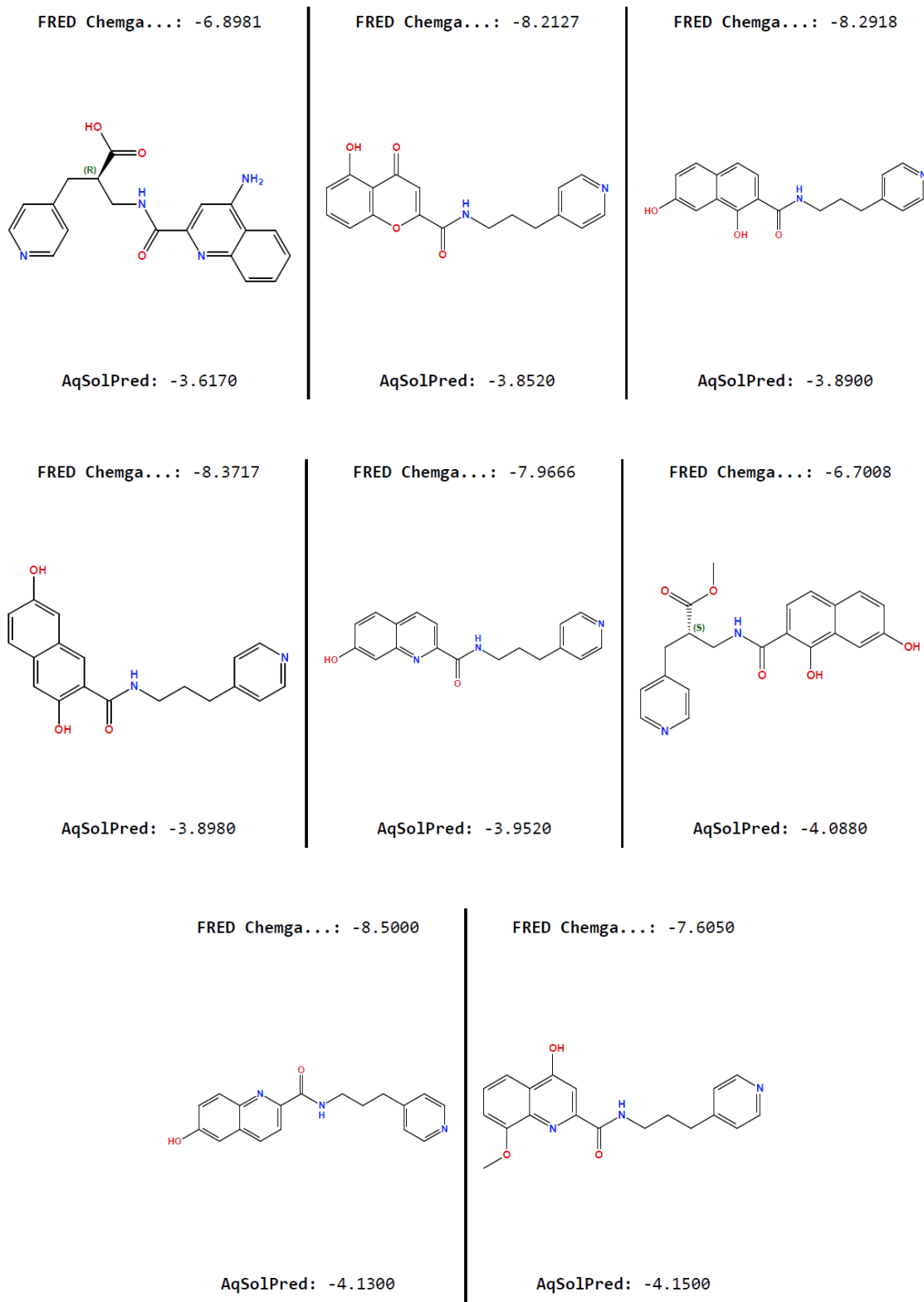


Figure 3.48. List of 19 analogues and the parent molecule CACHE3HI_1689_35 (first molecule in the list) selected for experimental testing. The molecules are sorted from best to worst AqSolPred logS scores. The FRED Chemgauss4 docking scores are also mentioned at the top of each molecule.

3.8. Discussion

The group IV viruses have +ssRNA and are divided into various families having varied levels of sequence conservation between them. For example, NSP12 of the *Coronaviridae* family and NSP9 of the *Arteriviridae* family have approximately 200–400 highly conserved residues in the NiRAN and RdRp domains (Xu et al., 2003; Snijder et al., 2016). Considering the order *Nidovirales*, the N-terminal of NSP13 is found to be the most conserved domain (Gorbalenya, 2001; Nga et al., 2011). All these studies suggest sequence conservation between various +ssRNA viruses. Considering the sequence conservation among other viral families within group IV, it is also important to also explore the structural similarities of various viral proteins. For example, protease (NS3/4A) of the hepatitis C virus shares structural similarities (substrate binding cleft, active site) with the protease (Mpro) of SC-2 (Bafna et al., 2020). Such studies can pave the way towards finding drug alternatives that can be repurposed for treating SC-2 infection. Bioinformatics studies such as sequence alignments and structural overlays of viral proteins of SC-2, betacoronaviruses or other group IV viruses, can help with identifying conserved residues. There is a need to perform a comprehensive sequence alignment study on these viruses; the conserved regions that have a role in protein activity could be targeted for further binding site analysis with suitable inhibitors.

The bioinformatics study was focused on SC-2 NSP3 and NSP13 protein sequence analysis and structural overlays with betacoronaviruses, group IV viruses and chikungunya virus (CHIKV) sequences. For both the proteins, NCBI and GISAID databases were used to download the SC-2 sequences. Statistical analysis showed that the NCBI dataset, our initial choice, had more submissions from the USA, which might lead to bias within our analysis. Therefore, we decided to include a new database in the study, GISAID, which contained sequence submissions from all over the world. Statistical analysis of the GISAID dataset showed a better sequence geographic distribution compared to the NCBI dataset, despite the fact that most submissions still came from the USA. The MSA of the SC-2 variants from both datasets showed conserved and intermediately conserved binding pockets of NSP13 and NSP3, respectively. It was observed that the mutations identified in the two sequence datasets had slightly different top 10 mutations, which points out the importance of sequence dataset size in order to identify more populated mutations. A few of the

mutations identified in the analysis led to a change in the amino acid property, which should be further explored to understand the biological implication of these mutations. Now considering the betacoronaviruses, the MSA of SC-1, SC-2, and MERS-CoV also showed high binding site sequence similarities for both helicase and Mac1 domain proteins. The sequence as well as structural comparison was also done considering human co-related proteins to Mac1 and helicase to rule out any chances of off-target binding of small molecules designed to target the viral proteins. It was found that the human RNA helicases do not share conserved sequence or structure with SC-2 NSP13. But on the other hand, the human PARP14 showed sequence similarity of ADP-r binding site and high overall structure similarity to SC-2 Mac1-NSP3. As a result, the CACHE organisation tested the successful hits from challenge #3 in a binding assay against PARP14 to rule of off-target binding. Despite having similarities between human counterparts and the SC-2 NSP3 Mac1 domain, it is important to note that there is low sequence similarity in regions adjacent to the ADP-r recognition motifs, which indicates that the likelihood of cross-reactivity between inhibitors of viral and human proteins is minimal (Rack et al., 2020). The SC-2 Mac1 which is a MacroD-type macrodomain found within NSP3 facilitates the removal of mono-ADP-ribose from modified proteins. It has conserved catalytic motifs (NAAN, GGG, GIF, VGP) and substrate-binding residues (D22, N40), which are also present in human MacroD1/2 (de-ADP-ribosylase). However, the overall sequence similarity is below 38%, suggesting the possibility for selective inhibition of the viral Mac1 domain, despite the few shared similarities (Rack et al., 2020). Our study focussed on the computational design and primary binding assays of SC-2 Mac1 inhibitors, it did not include biological investigations like assessing the impact of the designed inhibitors on viral replication, host ADP-ribose signalling, or any potential off-target activity against human de-ADP-ribosylases. But future biochemical and cellular studies will be beneficial in identifying antiviral efficacy of these inhibitors thus confirming their selectivity towards the viral Mac1 domain.

Upon comparing the group IV viral sequences and protein structures, it was found that CHIKV shared a high level of sequence as well as structural similarities with SC-2 NSP13 and Mac1-NSP3. Additional MSA on CHIKV sequences was also done to determine if the binding pockets of CHIKV nsP2 (helicase) and nsP3 (Mac1) were also conserved. It was found that the binding sites of the CHIKV protein sequences were highly conserved. But it is important to note that the MSA on CHIKV sequences consisted of 280 sequences, which is a small

dataset that can miss out on populated mutations. The binding sites' structural overlays of the helicase and Mac1 domain proteins between the two viruses showed high structural similarities. Though the RMSDs were calculated considering the AAs present in the binding sites, structurally there was good overlap between the proteins of the two viruses. As mentioned in Chapter I, there is no approved small molecule for the treatment of CHIKV infection, which falls under the epidemic and pandemic-prone disease category by the WHO alongside coronavirus infection. Since there are sequence and structural similarities between SC-2 and CHIKV helicase and Mac1 domain proteins, a pan antiviral could be developed for treating the two infections.

Due to the mandate of the CACHE challenge, this study was focused on NSP3 and NSP13. The output of the MSA gave insight into the level of conservation across these NSPs (discussed in Chapter II). There were highly conserved residues within the binding pockets of NSP3 (ADP-r binding pocket of the Mac1 domain) and NSP13 (RNA-binding tunnel). In the case of NSP3, there were known actives that bind to the Mac1 domain. The fastROCS volume query, pharmacophore query and docking protocols were validated due to the presence of these positive controls in pocket 1. It was observed that two X-ray structure PDB submissions (7KQP and 6WOJ) had different protein conformations due to alternative orientations of isoleucine 129. One conformation of isoleucine that had the side chains inwards resulted in a closed tunnel conformation (6WOJ), whereas if the side chains were pointing outwards, it resulted in an open tunnel conformation (7KQP). The two structures were included in the pocket selection process, i.e., pocket 5 (closed tunnel) and pocket 6 (open tunnel). In total, six pockets were selected with the placement of dummy atoms in the Mac1 domain of NSP3. In contrast, NSP13 had no X-ray crystal structures containing lead-like compounds co-crystallized in the RNA-binding tunnel; instead, only fragments were present. As there were no known actives for the binding pocket of NSP13, the VS protocols could not be validated. The VS workflow for NSP13 targeted ten pockets in total.

The VS protocol in both the studies (NSP3, NSP13) resulted in hits that had passed through three rounds of screening. For both the VS workflows, ~1.4 billion molecules were screened from the Enamine database. In the case of NSP3, the fastROCS screening resulted in a total of 450,000 hits for all six pockets. The output of pharmacophore searches and docking resulted in 8,050 and 7,296 hits, respectively. Whereas, for NSP13, the fastROCS screening resulted in a total of 6,000,000 hits for both the sites. The output of pharmacophore

searches and docking resulted in 151,327 and 135,648 hits, respectively. These hits had fulfilled both fastROCS and pharmacophore queries, and their best interaction poses were predicted by docking. The hits made interactions with protein residues that were conserved across the datasets.

For this study, only the FRED docking algorithm was considered; due to time constraints, other docking tools could not be explored to identify a consensus docking pose. In a separate project with collaborators, we utilised a consensus docking approach by using MOE, ClusPro and ZDOCK tools to predict a more realistic interaction pose between respiratory syncytial virus NS1 and human STAT1 proteins (Efstathiou et al., 2024) and found that all three tools predicted a common docked pose between both proteins (Figure 3.49.).

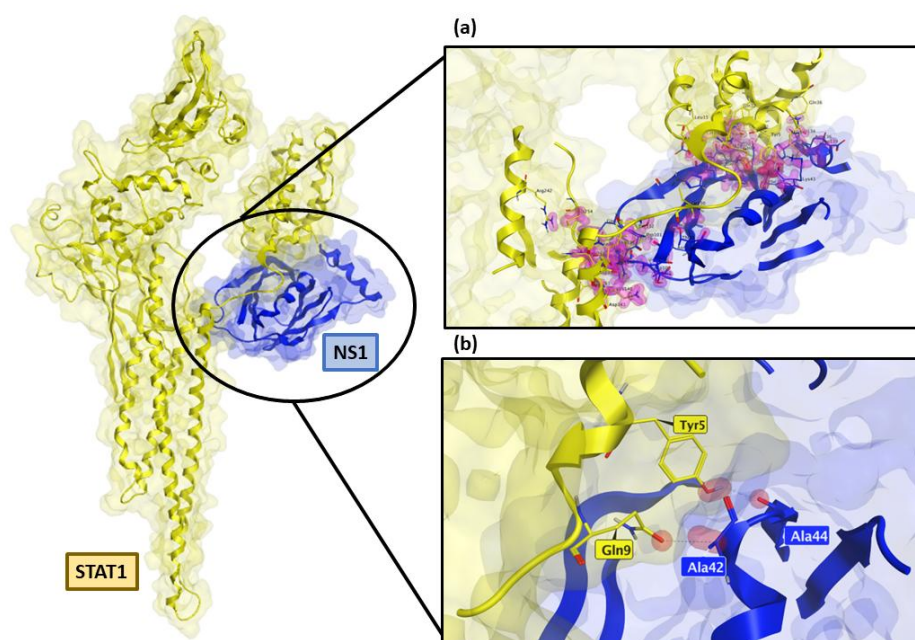


Figure 3.49. Consensus docked pose of RSV-NS1 (blue) and STAT1 (yellow). The contact amino acids are highlighted in red. Schematic taken from Efstathiou et al., 2024.

The experimental results by CACHE for the NSP13 challenge indicated binding of 5 of our hits in the primary SPR run. But compounds with $RU > 2.0$ were excluded, as they are indicative of non-specific binding. There was one hit (CACHE_1432_1) with $RU = 2.29$ (Figure 3.23.); even though it is above the RU threshold value decided by CACHE, it can still be intriguing to investigate its binding to helicase protein in dose-response SPR binding assays. The docked pose of CACHE_1432_1 at siteA2 (Enamine ID Z4414679777) had a Chemgauss4 score of -11.7 and made four hydrogen bond interactions with the protein

backbone (Figure 3.50.). From the 5 hits, structural analogues of CACHE_1432_28, which had K_d 551 μ M and % binding 268, can be retested since, although it had a poor K_d , it displayed good % binding, along with acceptable DLS aggregation and no solubility issues. The validation of the VS protocol was done with the data released by CACHE, consisting of active and inactive molecules from the challenge. These results helped in modifying the VS workflow and identifying that during pharmacophore screening, most of the actives were lost. Thus, the pharmacophore search query was modified, which helped in identifying a greater number of actives.

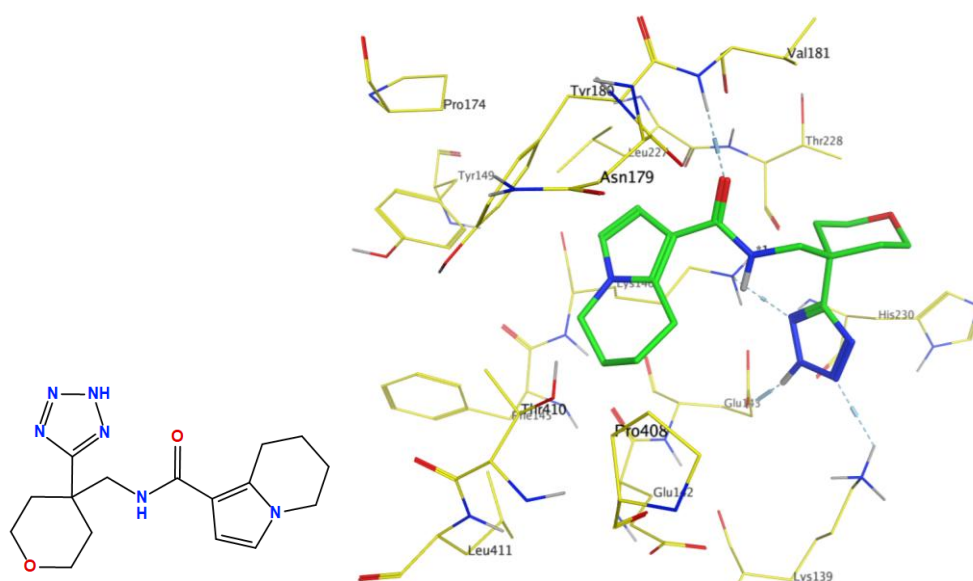


Figure 3.50. Structure and docked pose of CACHE_1432_1 (green) at siteA2 (yellow). Hydrogen bonds are depicted in dotted lines.

The experimental results by CACHE for the NSP3 challenge showed binding of 2 of our hits in the primary HTRF assay and subsequent dose-response SPR runs. The compound CACHE3HI_1689_15 had K_d values of 328 μ M and >1000 μ M at 2% and 4% DMSO concentrations, respectively. This compound was not selected for round 2 evaluation because it missed the K_d cut-off value. The compound CACHE3HI_1689_35 had K_d values of 4.84 μ M and 2.6 μ M at 2% and 4% DMSO concentrations, respectively, and was also not selected for round 2 evaluation because it missed the % binding cut-off value, noisy sensorgram in SPR assay and showed high solubility issues in DLS testing. which indicates a solubility issue. The validation of the VS protocol was redone with the data released by CACHE, consisting of active and inactive molecules from the challenge. This helped in modifying the VS workflow and identifying that at the pharmacophore screening stage,

most of the actives were lost. The pharmacophore search query was modified, which helped in identifying a greater number of actives. The Mac1 domain binds to ADP-r, and the molecules designed in this study act as competitive inhibitors not only for ADP-r but also for the actual substrate, which is protein-ADP-r. Since the active site of the SC-2 Mac1 domain was utilised to design these inhibitors, they will act as competitive inhibitors regardless of the presence of free ADP-r or protein-ADP-r. Future studies are necessary to validate the competitive inhibition of the designed inhibitors when both free ADP-r and protein-ADP-r are present.

The two molecules, CACHE3HI_1689_15 and CACHE3HI_1689_35, seemed promising but were not selected for further testing due to the cut-off values selected by CACHE. A separate hit to lead study was carried out with the aid of the Notre Dame-DCU Seed Grant in Precision Biomedical Technologies funding. The structural analogues of the two hits were searched in the Enamine REAL database, and the solubility (logS) was predicted with the AqSolPred ML model, as this model performed slightly better than MOE in predicting the solubilities of the molecules in the benchmarking study. The 19 selected analogues, along with the parent molecule CACHE3HI_1689_35, are currently being shipped to our collaborator Dr Choon Kim at Notre Dame University, USA. These analogues will be tested in binding assays (SPR) against the SC-2 Mac1-NSP3 as well as the CHIKV Mac1 domain. The experimental validation results are expected to be completed by the end of September 2025. Our collaborator Prof. Juan Hermoso at CSIC in Spain will conduct a confirmatory X-ray crystallography if the binding assays identify any hits.

3.9. Conclusions and future directions

- Multiple studies have highlighted that there are sequence similarities across coronaviruses as well as the group IV viruses (+ssRNA), which should enable the identification of conserved regions within the viral genome and target the conserved residues present in the protein binding sites.
- By utilising bioinformatics tools (MSA), conserved binding pockets in the SC-2 NSP13 (RNA binding site) and the SC-2 Mac1 domain of the NSP3 (ADP-r binding site) were determined across the betacoronaviruses and the SC-2 variants.

- High levels of sequence and structural conservation were established between group IV viruses, Chikungunya (CHIKV) and SARS-CoV-2, for helicase (NSP13) and Mac1 domain (NSP3) proteins. These conserved regions may represent novel starting points for designing small molecule interventions across both viruses. Their conservation points to their importance in the viral life cycle within which they may form key protein-protein interactions (Fayne, 2013) and interact with cofactors or other essential biological agents.
- The computational drug design approaches used for the CACHE challenges involved VS of the Enamine database to elucidate interactions between novel NSP binding sites and potential small molecule inhibitors. The hit lists obtained after the VS workflows were sent to CACHE, where the compounds were tested by performing binding assays.
- There were five interesting hits identified for challenge #2 (SC-2 NSP13) after the completion of binding assays but were not selected for further testing as they did not pass the cut-off criteria established by CACHE. Structural analogues of these hits and another molecule, CACHE_1432_1 (RU=2.29), can be retested.
- There were two interesting hits identified for challenge #3 (SC-2 NSP3) after the completion of binding assays but were not selected for further testing as they did not pass the cut-off values established by CACHE. A separate hit to lead study on these hits is being carried out with the aid of the Notre Dame-DCU Seed Grant in Precision Biomedical Technologies funding.
- The 19 selected analogues, along with the parent molecule CACHE3HI_1689_35, will be tested in binding assays against the SC-2 Mac1-NSP3 as well as the CHIKV Mac1 domain by Dr Choon Kim at Notre Dame University, USA. A confirmatory X-ray crystallography will be done by our collaborator Prof. Juan Hermoso at CSIC in Spain if the binding assays identify any hits.
- Since the binding site of Mac1 shares similarity with human PARP14, it is important to consider off-target implications of the identified small molecule inhibitors.

- Due to the conservation of these sites across betacoronaviruses (SC-2, SC-1, MERS-CoV) and CHIKV, it may be possible to discover pan-antiviral small molecule inhibitors offering protection from possible future outbreaks of related viruses.
- The hit molecules identified for SC-2 proteins, i.e., NSP13 and NSP3, can be used as starting points for creating more potent analogues. The CADD protocol developed in the project can be replicated to identify initial hit molecules for any protein of interest with the requisite data.

Chapter IV.

Molecular Dynamics simulations & Absolute Binding Free Energy calculations

Parts of this chapter have been published in:

Zhang Y, **Kandwal S**, Fayne D, Stevenson NJ. MERS-CoV-nsp5 expression in human epithelial BEAS 2b cells attenuates type I interferon production by inhibiting IRF3 nuclear translocation. Cell Mol Life Sci. **2024**, 81(1), 433.

4.1. Introduction

In this chapter, we will discuss the molecular dynamics (MD) simulation workflow developed for the two identified hits of SC-2 Mac1-NSP3. Furthermore, we will explore the absolute binding free energy (ABFE) workflow that has been implemented to predict the binding affinities of these hits in comparison to the experimental binding affinity.

4.2. Why MD is necessary?

In a traditional SBDD, molecular docking is used to predict an energetically favourable binding pose between the ligand and the protein binding site. It basically involves two steps: first, sampling the ligand poses in the binding site and second, utilising a scoring function on each pose. In the sampling stage, the search algorithms look for favourable conformations for both the protein binding site and the ligand, if using fully flexible docking. Whereas in the second stage these conformations are ranked using scoring functions (Agu et al., 2023). These scoring functions evaluate the ligands based on estimating how favourable the protein-ligand interactions are in a predicted pose. They are classified into forcefield-based, empirical function and knowledge-based scoring. Despite having the sampling component in the docking tools, the conformations of the protein and ligand are not fully explored since they possess a high level of flexibility and undergo conformational changes in solution. In the case of rigid-rigid docking, the docking algorithm only explores translations and rotations of the rigid ligand within the rigid protein binding pocket. This situation is where MD simulation helps in resolving the issue by modelling conformational changes of the protein-ligand complex in solution (Figure 4.1.).

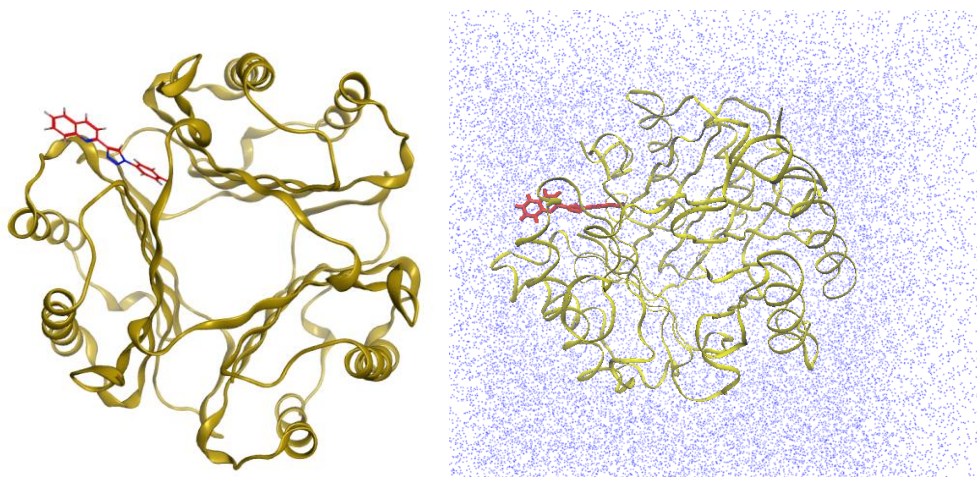


Figure 4.1. Example of protein-ligand (Macrophage migration inhibitory factor (MIF) protein iaryltriazole inhibitor 3b-180, PDB: 4WR8) docked pose (left) alongside its MD setup in solution (right). Ligand (red), protein (yellow), water (dark blue). The image was generated using MOE 2022.02 and VMD.

4.3. Molecular Dynamics simulations

In drug discovery, MD simulations are widely used to explore and understand interatomic interactions between different systems, e.g., protein-ligand complexes or protein-protein complexes, by predicting the position of each atom over time using classical mechanics. The first MD simulation work on proteins was published in the 1970s (McCammon et al., 1977). Richard Feynman, who received the Nobel Prize in physics in 1965, stated that 'all things are made of atoms, and that everything that living things do can be understood in terms of the jiggings and wiggings of atoms.' Much of the biophysics now focuses on exploring the jiggings and wiggings of atoms by using MD simulations. The basic principle behind MD simulation is the calculation of the forces exerted on each atom by all the other atoms, and by using Newton's laws of motion, the spatial position of each atom as a function of time is predicted. These forces are calculated continuously over a period of time, which are then used to update the positions and velocities of each atom. The output of this process is a trajectory, which can be viewed as a 3D movie showing the configuration of each atom over a time interval (Hollingsworth et al., 2018) (Figure 4.2.).

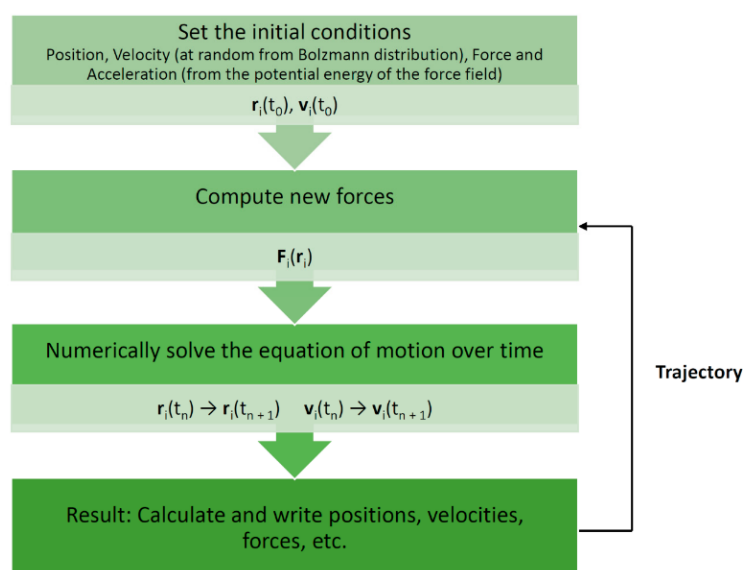


Figure 4.2. Steps involved in generating MD trajectory from initial positions followed by computing of forces where r-position; v-velocity; t-time; F-force. Schematic taken from (Salo-Ahen et al., 2021).

The forces acting on each atom are computed by a model or empirical potential energy, i.e., a force field, which is a function of atom positions. The force fields are mathematical models which consist of parameters derived from the results of quantum mechanical calculations and experimental measurements. The forces on atoms are thus calculated by considering these parameters; the forces arise from interactions between bonded and non-bonded atoms. The bonded interactions include bond stretch, bond angles and dihedrals. The covalent bonds and bond angles are represented as harmonic springs, while dihedral rotations around bonds are described using sinusoidal functions that capture the energy differences between staggered and eclipsed states. Non-bonded interactions are accounted through van der Waals forces, modelled with the Lennard–Jones potential, and electrostatic interactions, represented by Coulomb’s law (Durrant & McCammon, 2011) (Figure 4.3.).

$$E_{\text{total}} = \underbrace{\sum_{\text{bonds}} K_r (r - r_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)]}_{\text{Bonded}} + \underbrace{\sum_{i < j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]}_{\text{Non-bonded}}$$

Figure 4.3. Example of equation used to approximate the atomic forces by taking into account both bonded and non-bonded interactions. The chemical bonds and bond angles are represented using simple spring models. The dihedral (torsional) angles, which describe rotations around bonds, are represented by a sinusoidal potential for the identification of energy variation between eclipsed and staggered conformations. The non-bonded interactions include van der Waals forces, modelled by the Lennard-Jones potential, as well as electrostatic interactions, described by Coulomb's law. Schematic taken from (Durrant & McCammon, 2011)

4.4. CHARMM forcefield

The force fields used in MD simulations can be classified into additive and polarisable force fields. The additive forcefields like AMBER, CHARMM, OPLS, and GROMOS are used to calculate the electrostatic interactions by considering static or fixed partial atomic charges, which are not ideal to model environments where the charge distribution is expected to

change. To tackle this problem, polarisable force fields AMOEBA or CHARMM-Drude, which take into account the changes in charge distribution, are used (He et al., 2021).

The forcefield, Chemistry at HARvard Macromolecular Mechanics (CHARMM) (MacKerell et al., 1998; Huang et al., 2017), is an all-atom additive forcefield which has been in development since the early 80s. It is used to assign predefined parameters for proteins, nucleic acids, lipids and carbohydrates. In 2011, it was updated to CHARMM General FF (CGenFF) (Vanommeslaeghe et al., 2011), which calculates parameters of a wide range of chemical space. The most recent version of the force fields is CHARMM36m (Huang et al., 2017) for biomolecules and CGenFF version 4.6 for small molecules. While generating the parameters for the small molecules, CGenFF assigns penalty scores, which offer an estimate of how well the molecule is parametrised. The penalties lower than 10 indicate the parameters are fair; the penalties between 10 and 50 mean some basic validation is recommended; penalties higher than 50 indicate a poor parameter and mandate extensive validation and optimisation.

4.5. GROMACS

GROMACS (Pronk et al., 2013) is a versatile open-source MD simulation package for simulating proteins, lipids and nucleic acids using Newtonian laws and equations of motion. To model the interactions between protein-ligand complex, the parameters are computed for the complex using forcefields.

For a system N interacting atoms:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i, \quad i = 1 \dots N \quad (8)$$

Where, the forces are the negative derivatives of a potential function $V(r_1, r_2, \dots, r_N)$:

$$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i} \quad (9)$$

The overall force on any atom F_i is then computed by calculating forces between non bonded, bonded atom pairs and restraining and/or external forces. The potential and kinetic energies along with pressure tensor may also be computed.

$$\mathbf{F}_i = \sum_j \mathbf{F}_{ij} \quad (10)$$

Using the forces, the movement of the atoms are simulated by solving Newton's equations of motion.

$$\frac{d^2\mathbf{r}_i}{dt^2} = \frac{\mathbf{F}_i}{m_i} \quad \text{or} \quad \frac{d\mathbf{r}_i}{dt} = \mathbf{v}_i \quad \frac{d\mathbf{v}_i}{dt} = \frac{\mathbf{F}_i}{m_i} \quad (11)$$

In an MD setup, the complex is then placed in a water box with periodic boundary conditions (PBC). This step is done to avoid problems with boundary effects, which are caused by finite size; thus, the system is considered more like an infinite one. Additionally, to keep the atoms in a fixed reference position, positional restraints are added, particularly during the equilibration stage, to avoid drastic rearrangements. The MD simulation workflow is divided into four stages. The first stage is the energy minimisation stage, where the system is relaxed to sort out any undesirable starting structures or steric clashes. The second stage is the NVT equilibration ensemble stage, where the number of particles (N), volume (V) and temperature (T) are constant. The third stage is the NPT ensemble equilibration stage, where N, pressure (P) and T are constant. The last stage is the MD production stage, where the simulation runs to generate the MD trajectory under the NPT ensemble but without the positional restraints (Figure 4.4.).

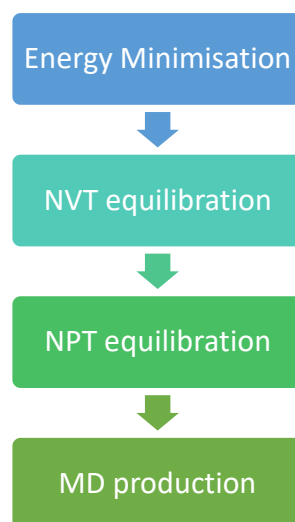


Figure 4.4. Flow chart of different stages involved in MD simulation. There is one energy minimisation step, followed by two equilibration steps (NVT and NPT), and lastly, the MD production step

The main files of GROMACS are the topology (.top), which contains a complete description of all the interactions in the protein; the molecule structure file (.gro, .pdb), which contains

xyz co-ordinates; the MD parameter file (.mdp), which contains information about the four stages of MD simulation; the index file (.ndx), which helps in specifying the action on groups of atoms; the run input file (.tpr), which contains all the information that is needed to start a simulation with GROMACS; and the trajectory file (.trr, .tng, or .xtc), which is the output file containing the trajectory. A typical GROMACS workflow is shown in Figure 4.5., where the input eiwit.pdb is used for generating a GROMACS topology file, defining a solvent box, solvating the protein, and running the four stages of the MD simulation. The trajectory files are then analysed.

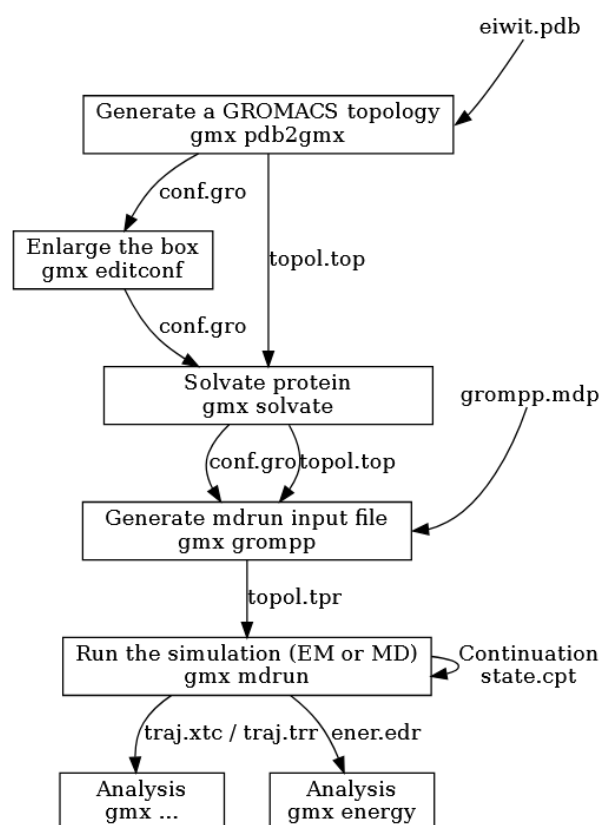
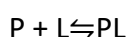


Figure 4.5. A typical MD workflow using GROMACS. The input file for MD is a PDB file which is converted to a gmx file followed by the generation of the topology file. The protein is solvated and grompp is used to generate the tpr file. The simulations are then run, starting with EN, NVT, NPT, and MD. Schematic taken from GROMACS documentation (<https://manual.gromacs.org/current/manual-2025.2.pdf>).

4.6. Absolute Binding Free Energy calculations (ABFE)

The accurate binding affinity prediction of a drug to its target protein has been one of the main objectives of CADD. The best and the most accurate way is alchemical free energy

calculations which relies on all-atom MD simulations in explicit solvent. It depends on statistical mechanics principles and takes into consideration the entropic and solvent effects. The calculations depend on a thermodynamic cycle that incorporates a series of non-physical intermediate states (alchemical). The energy difference between the two end states is obtained as the sum of the free energy differences among all pairs of alchemical intermediate states (Shirts et al., 2007). The ABFE calculations are based on such an alchemical cycle in which the ligand is non-physically “removed” from the solution and “inserted” into the protein’s binding site (Aldeghi et al., 2018). The binding of a ligand (L) to protein (P) can be described from the following chemical reaction:



The binding constant K_b^o at equilibrium is defined by the ratio of the product [PL] and the reactants concentrations [L][P] in solution, where c^o is the standard state concentration:

$$K_b^o = c^o \frac{[PL]}{[L][P]} \quad (12)$$

In case of binding event of a single ligand K_b^o is associated to the binding free energy by:

$$\Delta G_b^o = -k_B T \ln K_b^o \quad (13)$$

Where k_B is the Boltzmann constant and T is the temperature.

The best method to estimate free energy differences is based on perturbation theory, which is explained for a formula derived by Zwanzig (Zwanzig, 1954). The equation shows that by taking the logarithm of the average exponential of the potential energy difference between two states weighted according to the Boltzmann distribution, the free energy difference can be estimated. The formula applies to the forward transformation 0→1 and backward transformation 1→0 to calculate free energy difference considering “0” and “1” are thermodynamic states.

$$\Delta G_{0,1} = -k_B T \ln \left\langle e^{-\beta(U_1(x) - U_0(x))} \right\rangle_0 \quad (14)$$

To obtain a reliable estimate of free energy difference using the perturbation method several intermediate states are required. Considering protein-ligand complex, the two end states bound (protein-ligand complex) and unbound (ligand in solvent) states are not

sufficient i.e., a pathway of intermediates states is necessary. Hence, a thermodynamic cycle that connects the bound and unbound states with several intermediates states is utilised for estimating the free energy difference for a protein-ligand complex. With the help of MD simulation these N intermediates states, also referred to as windows, are independently simulated. The free energy difference between each state i and its successor $i + 1$ is calculated and the binding free energy is recovered as the sum of all $\Delta G_{i,i+1}$.

$$\Delta G_b = \sum_{i=0}^{i=N-2} \Delta G_{i,i+1} \quad (15)$$

In the case of the protein-ligand complex absolute binding free energy cycle, there are a total of six states involving the two physically meaningful end states, i.e., bound and unbound states, and four alchemical intermediate states. In the intermediate states, the ligand is removed, or "decoupled," from the environment by turning off the intermolecular, electrostatic (elec), and van der Waals (vdW) interactions (Figure 4.6.). The intramolecular interactions of the ligand are also removed, as defined by the term "annihilated." In the unbound state (state A), only the ligand is simulated in a solvent box (orange), where all the intermolecular interactions (elec+vdw) are present. This state is then transformed to state B, where the ligand intermolecular interactions in solvent are removed (white) through several nonphysical states that are simulated individually. In state C, the ligand is restrained while the intermolecular interactions in the solvent are still removed (paperclip). The restraints are applied to the ligand to keep it within the binding pocket while perturbing the intermolecular interactions. State D also represents the restrained state, where the ligand is restrained within the binding pocket while the intermolecular interactions are completely removed. The intermolecular interactions are turned back on while restraining the ligand in state E (Aldeghi et al., 2018). Finally, the restraints are removed in state F, which is the end state, i.e., the bound state. The binding free energy is recovered by the sum of bound and unbound states free energies:

$$\Delta G_b^0 = -\Delta G_{elec+vdw+restr}^{prot} + \Delta G_{elec+vdw}^{solv} + \Delta G_{restr_on}^{solv} \quad (16)$$

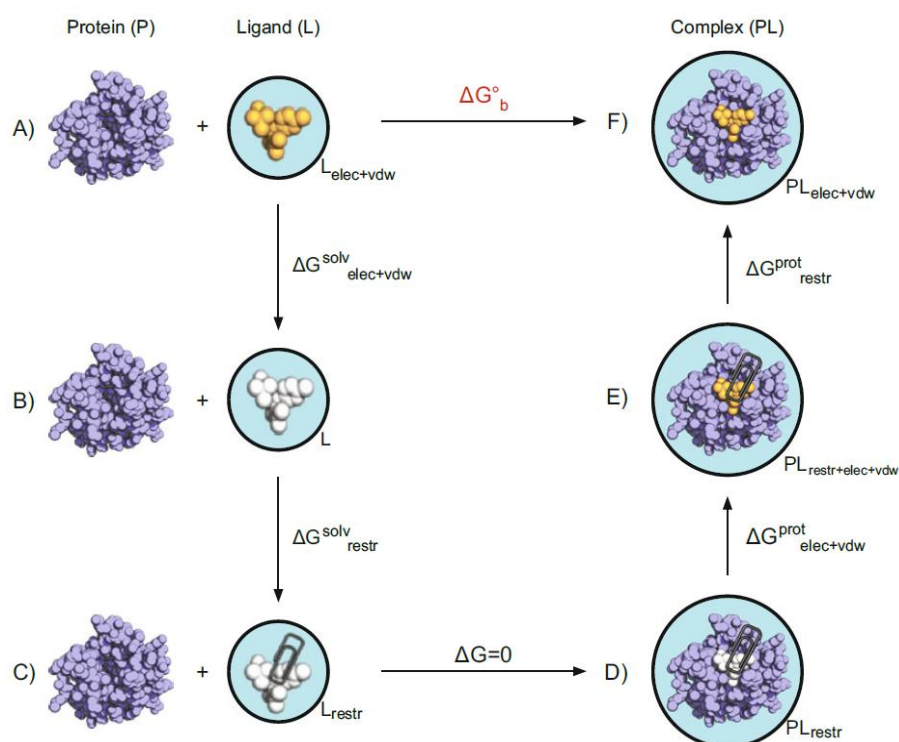


Figure 4.6. Non-physical thermodynamic cycle used to obtain the absolute binding free energies. The ABFE simulations include two states bound (complex-PL in solvent) and unbound (ligand-L in solvent). The protein is represented in purple, the solvent is represented in blue, the interacting ligand is represented in yellow, the non-interacting ligand is represented in white; and the restraint is represented as a paperclip. Schematic taken from (Aldeghi et al., 2016).

From the above intermediate states, different conformations are sampled from each individual window. In the different states, restraints (paperclip) were added to prevent the ligand from leaving the protein binding pocket while perturbing (slowly turning off) the intermolecular interactions. This is done to ensure that the conformations sampled during the simulations correspond to the bound state. The harmonic restraints include one distance, two angles, and three dihedrals which are applied between three atoms of the ligand and three atoms of the protein (Figure 4.7.). These are a set of Boresch restraints (Boresch et al., 2003) which allow the ligand to be kept in a specific orientation in the binding pocket and also provide an analytical solution without running the simulation. But a state with restraints must be estimated by running simulations across the intermediate states. The $\Delta G_{restr_on}^{solv}$ can be calculated analytically by the equation which includes the standard state correction:

$$\Delta G_{\text{restr_on}}^{\text{solv}} - k_B T \ln \frac{V}{V^0} = k_B T \ln \left[\frac{8\pi^2 V^0}{r_{aA}^2 \cdot \sin \theta_a \cdot \sin \theta_A} \frac{(k_{r_{aA}} k_{\theta_a} k_{\theta_A} k_{\phi_{ba}} k_{\phi_{aA}} k_{\phi_{AB}})^{\frac{1}{2}}}{(2\pi k_B T)^2} \right] \quad (17)$$

The standard volume V^0 is taken as $1660 \text{ \AA}^3$. The parameter r_{aA} represents the reference value for the distance restraint, while θ_a and θ_A are the reference values for the two angle restraints. The force constants k correspond to the harmonic restraints applied to one distance (r_{aA}), two angles (θ_a and θ_A) and three dihedrals ($\phi_{ba}, \phi_{aA}, \phi_{AB}$).

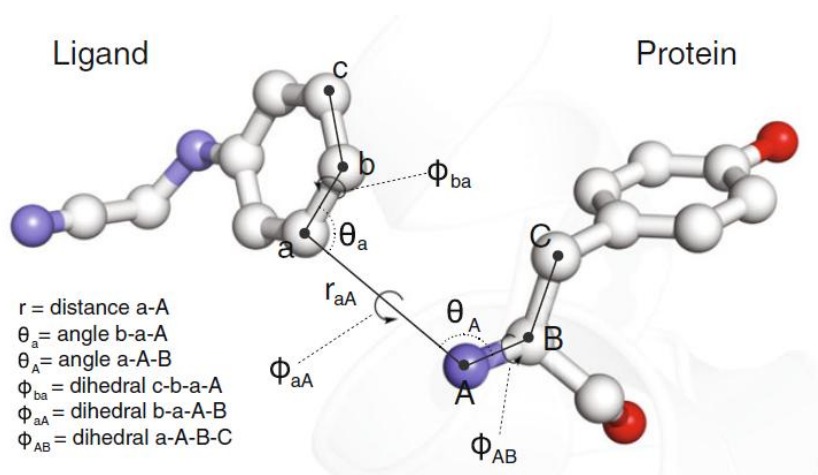


Figure 4.7. Set of Boresch restraints used for ABFE calculations where atoms a, b and c belong to the ligand and atoms A, B and C belong to the protein. Schematic taken from (Aldeghi et al., 2018).

The decoupling of the ligand across the different intermediate states, also referred to as λ windows, is facilitated by the decoupling parameter λ , which defines the thermodynamic state. In each λ window, the parameter can take any value between 0 (coupled) and 1 (decoupled) and is used to scale the coulombic charges, Lennard–Jones parameters and restraint force constants. The three transformations are carried out separately with distinct coupling parameters λ_{coul} , λ_{vdw} and λ_{restr} . In Figure 4.8., the coupling parameters λ_{coul} , λ_{vdw} and λ_{restr} are shown, which are scaled between 0 (coupled) and 1 (decoupled). In the coupled state the ligand is unrestrained, and the intermolecular interactions are kept on. Whereas in a decoupled state, the ligand is restrained, and the intermolecular interactions are turned off. During the alchemical transformation, the system experiences a change in energy, which can be plotted against the λ window. For Columbic interactions decoupling, there is a gradual drop in the energy, whereas for the vdW decoupling, there

is a gradual drop, followed by a large drop in the vdW curve because turning off steric (Lennard–Jones) interactions causes atoms to overlap unnaturally, releasing a big energy change, but later it reaches zero when the vdW interactions are completely removed (Figure 4.9.).

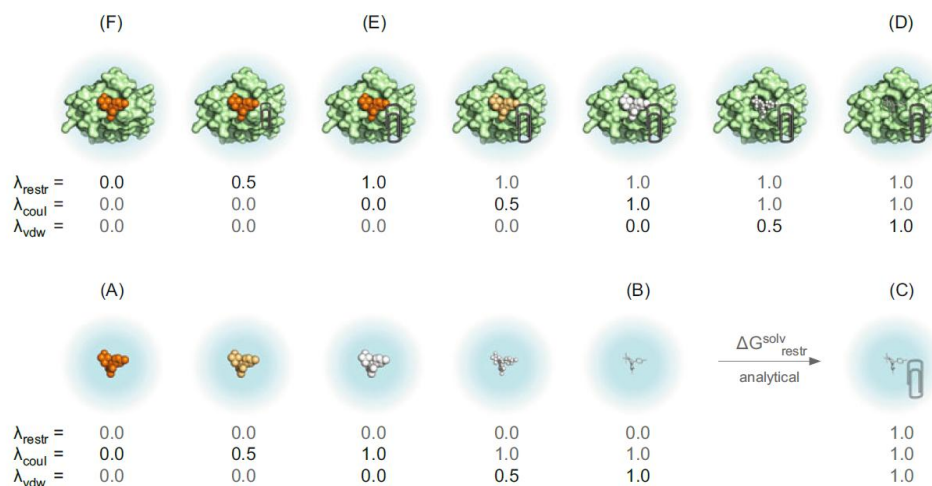


Figure 4.8. ABFE alchemical pathway and states controlled by the λ parameters. States A, B and C show the decoupling of ligand in solvent and states D, E and F show the decoupling of ligand from the binding pocket of the protein. The protein is represented in green, the solvent is represented in blue, the fully interacting ligand ($\lambda=0$) is represented in orange, the intermediate interacting ligand is represented in yellow/white, the fully non-interacting ligand is represented in grey ($\lambda=1$) and the restraint is represented as a paperclip. Schematic taken from (Aldeghi et al., 2018).

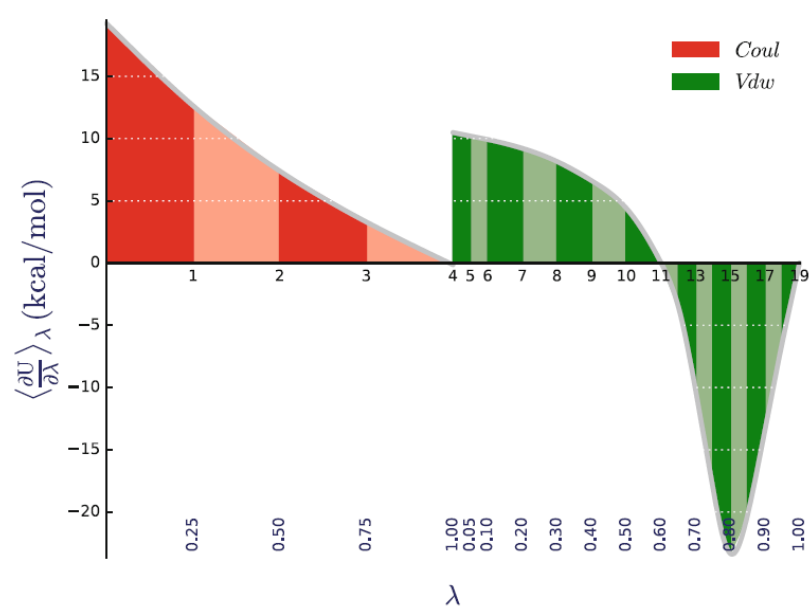


Figure 4.9. Example of plot showing change in energy during the decoupling of ligand. The red area represents coulombic interactions, the green area represents Vdw interactions. The x-axis shows the λ windows and the y-axis shows the change in energy (kcal/mol). Schematic taken from (Aldeghi et al., 2018).

4.7. Materials and Methods

4.7.1. Materials

- **Software:** GROMACS version 2020.7 (Pronk et al., 2013), Notepad++ version 8.4.6, Microsoft excel, Windows Subsystem for Linux (WSL), python version 3.8.10, hbmap2grace (<https://github.com/LMDM/hbmap2grace>), XMGRACE (Turner et al., 2005), MOE 2022.02, VMD 1.9.4a53 (Humphrey et al., 1996), Avagadro 1.1.1. (Hanwell et al., 2012), MDRestrainsGenerator (Alibay et al., 2022) (<https://github.com/Alibay/MDRestrainsGenerator>), MDAnalysis (Michaud-Agrawal et al., 2011)
- **Scripts:** cgenff_charmm2gmx_py3_nx2.py, hbmap_plot.py, sort.pl, data_collection.sh, em.sh, nvt.sh, npt.sh, md.sh, npt_second.sh, run_all.sh, mk_mdp.sh
- **GROMACS MDP files:** ions.mdp, em.mdp, nvt.mdp, npt.mdp, md.mdp
- **PDB files:** 5R82, 4WR8, 5SPT, 5SOJ, docked poses of CACHE3HI_1689_35 and CACHE3HI_1689_15
- **Operating systems:** Microsoft Windows 11 Education, Linux OS
- **System:** Dell Precision 3650 Tower, Processor- 11th Gen Intel(R) Core™ i9-11900 @ 2.50GHz, 2496 Mhz, 8 Core(s), 16 Logical Processors with NVIDIA graphic card
- **Servers:** **Vimes-** AMD Ryzen Threadripper PRO 7975WX 32-Cores 32-bit 64-bit with NVIDIA GeForce RTX 4090, **Ook-** AMD EPYC 7552 48-Cores Processor 32-bit 64-bit with four Tesla T4 graphics card, **Rincewind-** AMD Ryzen Threadripper PRO 5965WX 24-Cores CPU 32-bit 64-bit with NVIDIA GeForce RTX 4060
- **Forcefield files:**
 1. CHARMM36m
(https://mackerell.umaryland.edu/download.php?filename=CHARMM_ff_params_files/charmm36-jul2022.ff.tgz)
 2. CGenFF version 4.6 (<https://cgenff.com/>)

4.7.2. MD simulations

(A) Benchmarking of MD runs

To investigate the time taken by MDG servers (Ook, Rincewind and Vimes) to complete a full 100 ns MD simulation on the MIF protein PDB: 4WR8, benchmarking of the runs was performed using 32 CPU threads. The protein-ligand complex was quick prepped using MOE and the MD was set up as mentioned in (C) part of section 4.7.2., except the system was neutralised by the addition of 3 Cl⁻ ions. The time taken by each server to complete an MD run was noted, and a graph was plotted on MS Excel.

(B) Positive control MD runs on SC-2 Mac1-NSP3

The two PDBs, 5SPT and 5SOJ (Gahbauer et al., 2023), of SC-2 Mac1-NSP3 had lead-like molecules co-crystallised in the binding site with K_i 52.3 uM and 78.6 uM, respectively. The protein structures were prepared using the MOE 2022.02 quick prep option, with the default Amber10:EHT force field to consider explicit hydrogen atoms and tautomeric states and resolve any breaks in protein structure prior to conducting restrained all-atom molecular mechanics minimisation and electrostatics calculations. The MD setup was the same as mentioned in (C) part of section 4.7.2., except for the production runs, 5SPT was ran for 50 ns and 5SOJ was ran for a 100 ns. The MD trajectories were analysed and RMSDs of the protein chains and the ligands from their starting structure were calculated.

(C) MD on CACHE SC-2 Mac1-NSP3 hits

The docked poses (from Chapter III) of the two Mac1 domain NSP3 hits, CACHE3HI_1689_35 and CACHE3HI_1689_15, were used to carry out the MD simulations, which were performed using GROMACS 2020.7. The PDB files were converted to GROMACS-compatible topology files. The protein chains were parameterised using the CHARMM36m force field, whereas the small molecules (ligands) were parameterised using the CGenFF server. Before ligand parameterisation, the ligands were saved as .mol2 files using Avogadro, and their bonds were listed in ascending order using the sort.pl script. After ligand parameterisation with CGenFF, the files were converted to GROMACS-compatible files by the cgenff_charmm2gmx_py3_nx2.py script. For each simulation, the protein-ligand complex was centred in a cubic water box with an edge distance of 20 Å (2 nanometre padding from each side). The box was solvated using the Transferable

Intermolecular Potential 3 Points (TIP3P) water model (Jorgensen et al., 1983). The simulations were conducted within periodic boundary conditions. Since the two systems were neutral, no ions were added. Energy minimisation (em) was performed using the steepest descent algorithm in 5,000,000 steps. The minimised system was then equilibrated with the NVT ensemble and subjected to a 1 nanosecond (ns) run with position restraints under the leapfrog integrator (Van Gunsteren & Berendsen, 1988) and V-rescale temperature coupling (Bussi et al., 2007). It was followed by another equilibration run with the NPT ensemble at a constant temperature of 300 K and constant pressure of 1 bar for 5 ns with position restraints under the leapfrog integrator, V-rescale temperature coupling and Berendsen pressure coupling (Berendsen et al., 1984). The MD production runs for each pose were set to 100 ns, with all restraints removed, V-rescale temperature coupling and Parrinello-Rahman pressure coupling (Parrinello & Rahman, 1981) (Figure 4.10.). Throughout equilibration (NVT and NPT) and production MD, all bonds involving hydrogens were constrained using the Linear Constraint Solver (LINCS) algorithm (Hess et al., 1997). The parameters of ions.mdp, em.mdp, nvt.mdp, npt.mdp and md.mdp are detailed in **Appendix 6.4**. To investigate the stability of the pose an extra 500 ns run was performed for CACHE3HI_1689_35. The GROMACS in-built tools were used to calculate the number of hydrogen bonds, the RMSD of the protein and the ligand from its starting structure (equation 18), the total potential energy, temperature, pressure, root mean square fluctuations (RMSF) of AA residues (equation 19), density, radius of gyration (Rg) (equation 20), and clustering analysis to output the most representative frame from the MD simulation.

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^N m_i |r_i(t) - r_i(t_0)|^2}{\sum_{i=1}^N m_i}} \quad (18)$$

Where m_i is the mass of atom i , $r_i(t)$ is the position of atom i at time t , $r_i(t_0)$ reference position and N is the total number of atoms.

$$\text{RMSF}_i = \sqrt{\frac{1}{T - T_0} \sum_{t=T_0}^T (r_i(t) - \langle r_i \rangle)^2} \quad (19)$$

Where $r_i(t)$ is the position of atom i at time t , $\langle r_i \rangle$ is the average position of atom i over the trajectory, T_0 is the starting frame and T is the ending frame.

$$R_g = \sqrt{\frac{\sum_{i=1}^N m_i |r_i - r_{COM}|^2}{\sum_{i=1}^N m_i}} \quad (20)$$

Where m_i is the mass of atom i , r_i is the position of atom i , r_{COM} is the center of mass of the system and N is the total number of atoms.

The output files were visualised using XMGRACE software, and the most representative frame was structurally overlaid on the predicted docked pose using MOE to calculate the RMSD. MDAnalysis tool was also used to generate the RMSD plots of the protein and ligand. The hbmap2grace software was used to convert the hydrogen bond log into a text file that lists the number of each hydrogen bond interaction with specific atoms of the small molecule and the protein. The *in-house* hbmap_plot.py Python script was used to generate the hydrogen bond graphs. The MD trajectories were visualised using VMD.

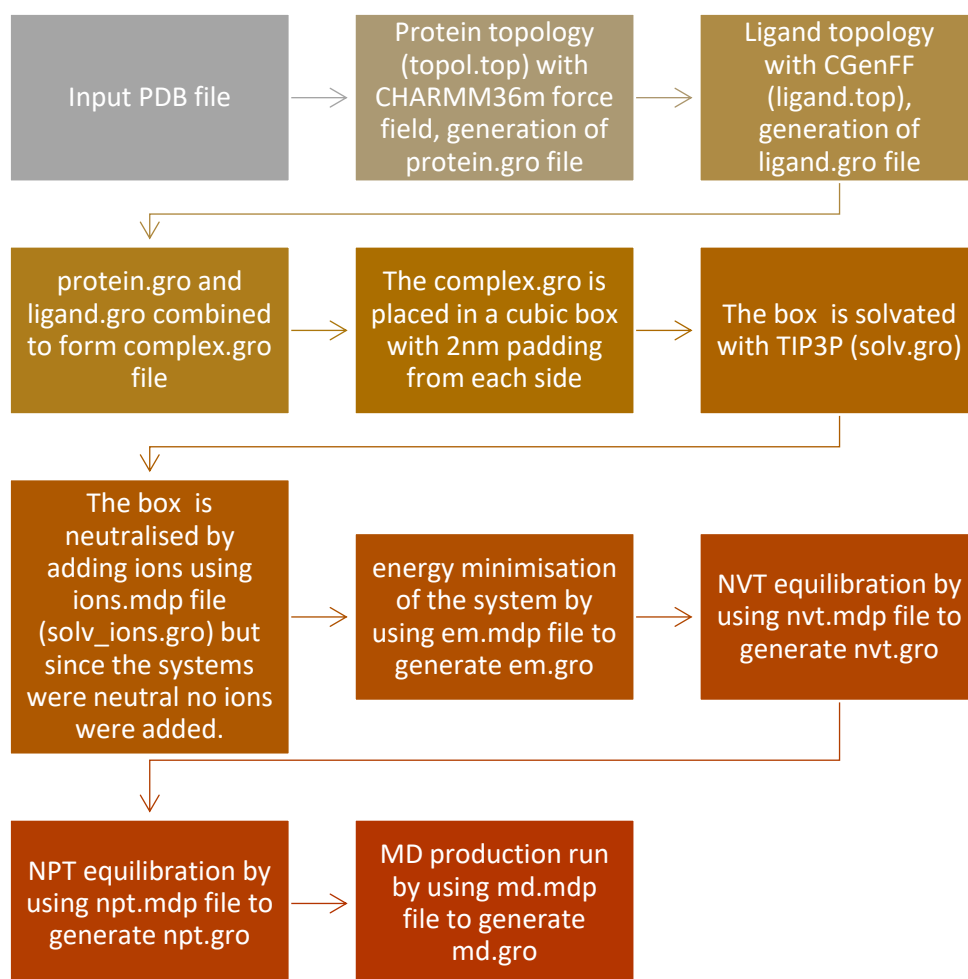


Figure 4.10. GROMACS MD simulation workflow. The input PDB file is converted to GROMACS-compatible files (top and gro). The protein-ligand complex is then solvated in a cubic box, followed by the addition of ions to neutralise the system's charge. The system is then subjected to energy minimisation, NVT/NPT equilibration and MD production.

4.7.3. ABFE calculations methods

(A) Benchmarking of ABFE protocol on SC-2 M_{pro} (NSP5)

The ABFE protocol was carried out with GROMACS, and to validate its capacity to run free energy calculations, an already existing ABFE workflow with GROMACS by Ghahremanpour et al., 2023, was replicated. The calculations were performed on SC-2 main protease (NSP5) (PDB: 5R82), which is in complex with Z219104216 (experimental $\Delta G = -6.05$ kcal/mol). Both protein and ligand were parameterised by the Optimized Potentials for Liquid Simulations (OPLS) force field (Robertson et al., 2015), and the input files for ABFE were provided. The calculations were divided into bound (protein-ligand complex) and unbound

states (ligand in solvent). In the bound state the decoupling of the ligand was done by gradually switching off the electrostatic interactions with a total of 11 λ windows, switching off the vdw interactions with a total of 21 λ windows, and the contribution of restraints was estimated by running 12 λ windows. Similarly, for the unbound state, the decoupling of the ligand was done by gradually switching off the electrostatic interactions with a total of 11 λ windows and switching off the vdw interactions with a total of 21 λ windows. Each λ window was an individual MD simulation with 5000 steps of steepest descent energy minimisation, 500 picoseconds (ps) of NVT equilibration, 1 ns of NPT equilibration and 5ns of MD production run. The free energies for bound and unbound states were computed individually using the GROMACS bar (gmx bar) module, which implements the Bennett Acceptance Ratio (BAR) method to estimate free-energy differences (in kJ/mol) between two neighbouring states. The final binding free energy was obtained by computing the above values with equation 16.

(B) ABFE protocol on SC-2 Mac1-NSP3

The ABFE workflow was carried on **two positive controls 5SPT and 5SOJ** of SC-2 Mac1-NSP3 which had lead-like molecules co-crystallised in the binding site and on **CACHE3HI_1689_35** (K_d 4.8 μ M). The K_i and K_d values were converted to ΔG by the following equations:

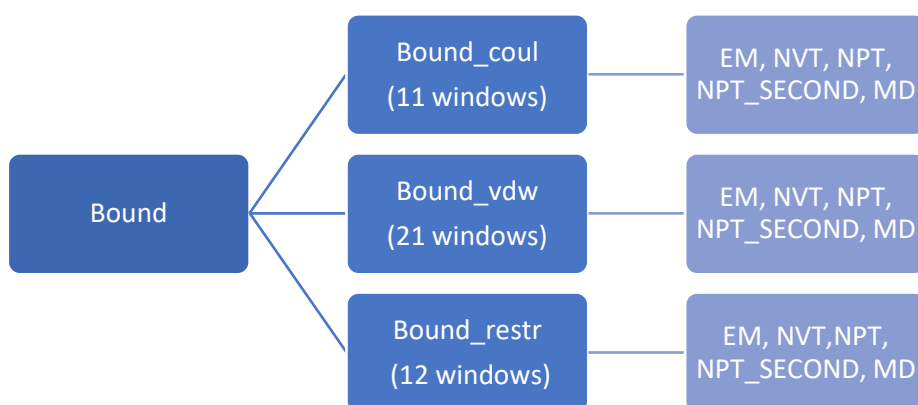
$$\Delta G = -RT \ln(K_i) \quad (21)$$

$$\Delta G = -RT \ln(K_d) \quad (22)$$

Where, T is the absolute temperature (298.15 K) and R is the universal gas constant (0.008314 kJ mol⁻¹ K⁻¹). The md.gro, index.ndx, ligand.top, and protein.top files from section 4.2.2.1. (A, B) were used as input files for the bound state simulations. The ligand in solvent files was generated by solvating the ligand in a cubic box with an edge distance of 25 Å (2.5 nm padding from each side). The system was energy minimised and equilibrated, followed by a 100 ns MD production run. The md.gro and ligand.top files were used as inputs for the unbound state simulations. To generate the Boresch restraints, MDRestraintsGenerator was used. The calculations were divided into bound (protein-ligand complex) and unbound states (ligand in solvent). In the bound state the decoupling of the ligand was done by gradually switching off the electrostatic interactions with a total of 11 λ windows, switching off the vdw interactions with a total of 21 λ windows, and the

contribution of restraints was estimated by running 12 λ windows. Similarly, for the unbound state, the decoupling of the ligand was done by gradually switching off the electrostatic interactions with a total of 11 λ windows and switching off the vdw interactions with a total of 21 λ windows. Each λ window was an individual MD simulation with 50,000,000 steps of steepest descent energy minimisation, 500 ps of NVT equilibration, 1 ns of NPT equilibration (with position restraints) at a constant temperature of 310 K and constant pressure of 1 bar, a second round of 1 ns of NPT equilibration (without position restraints) and 5 ns of MD production run (Figure 4.11.). The free energies for bound and unbound states were computed individually using the GROMACS bar (gmx) module to estimate free-energy differences (in kJ/mol) between two neighbouring states. The final binding free energy was obtained by computing the above values with equation 16. The λ window spacings for both bound and unbound states are mentioned in **Appendix 6.5**. The runs of each state were automated using bash scripts.

(a)



(b)

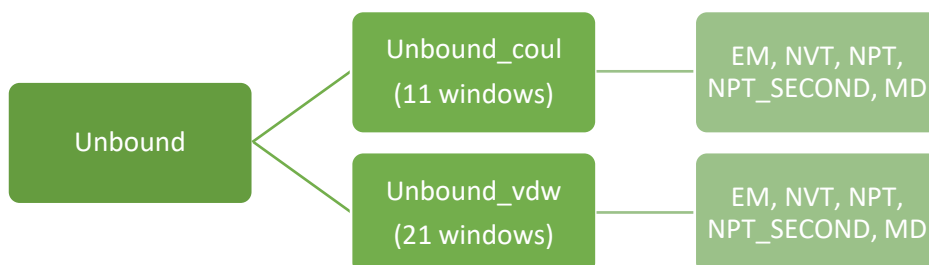


Figure 4.11. (a) Bound ABFE simulation workflow (blue) comprising bound_coul, bound_vdw and bound_restr decoupling states. (b) Unbound ABFE simulation workflow (green) comprising unbound_coul and unbound_vdw decoupling states. Each state undergoes MD simulations according to the λ window spacings, which include energy minimisation (EM), two stages of position-restrained equilibration (NVT and NPT), one stage without position-restrained equilibration (NPT_SECOND), and finally the MD production stage.

4.8. Results

4.8.1. Benchmarking of MD runs

The benchmarking study to investigate the time taken by MDG servers (Ook, Rincewind and Vimes) to complete a full 100 ns MD simulation on the MIF protein PDB: 4WR8 using 32 CPU threads showed that Vimes took the least amount of time (4 hours, 18 min), followed by Rincewind and Ook. With Ook runs, it was observed that when the number of GPUs was increased from 1 to 4, the time taken to complete the MD run decreased, but the runs with 2 and 3 GPUs show inconsistencies which can be due to GPU–GPU communication overhead and suboptimal domain decomposition, also reported in a study by Kutzner et al. (2015) (Figure 4.12). As Vimes took the least amount of time to complete MD runs, it was further used for ABFE calculations.

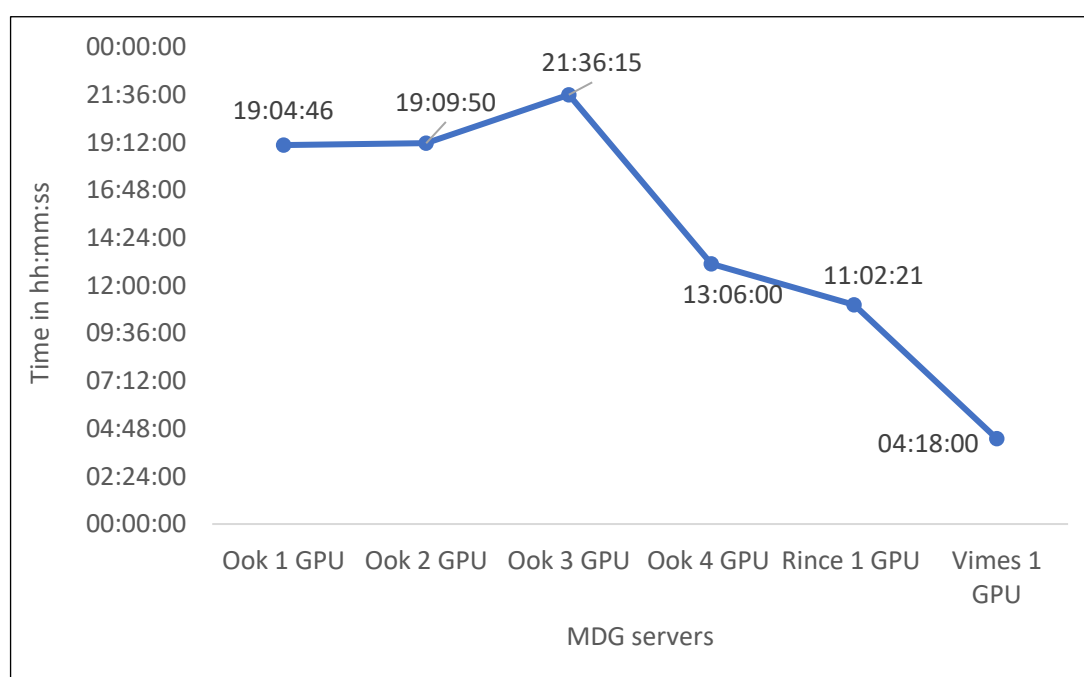
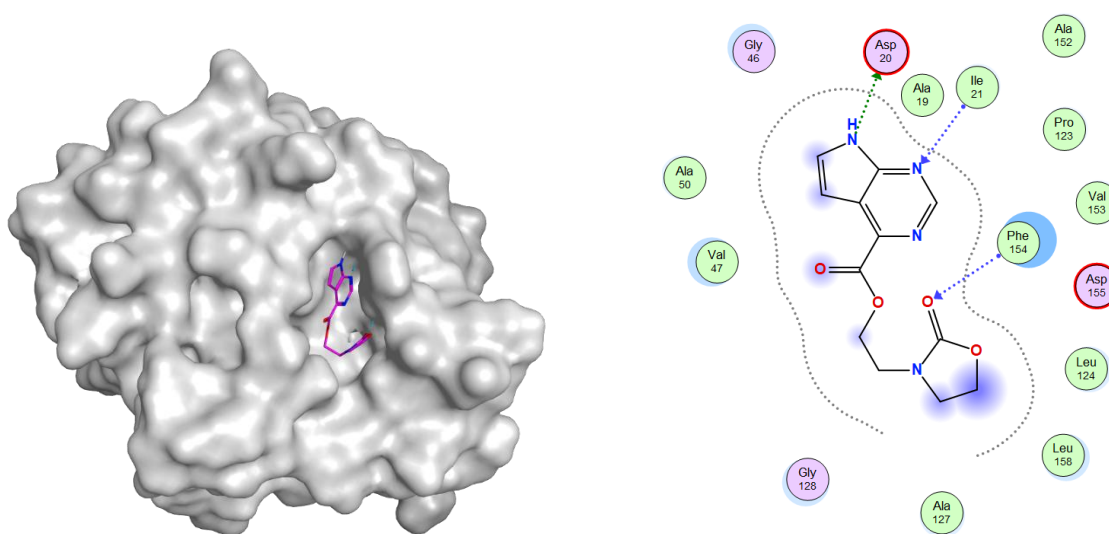


Figure 4.12. Benchmarking study plot between the MDG servers and time taken to complete a 100 ns MD simulation. The x-axis represents the MDG servers and the y-axis represents the time taken to complete the MD simulation (hh:mm:ss).

4.8.2. Positive control MD runs on SC-2 Mac1-NSP3

The two PDBs, 5SPT and 5SOJ had co-crystallised lead like molecules within the binding pocket of SC-2 Mac1-NSP3 which were selected as positive controls for the MD simulations. The ligand parameterisation with CGenFF resulted in a parameter penalty of 139 and a charge penalty of 68.1 for the ligand S6U (5SPT). Whereas the ligand RWQ (5SOJ) had parameter and charge penalties of 55.1 and 45.9, respectively. The RMSD plot of the protein-S6U complex (5SPT) between the starting pose and the MD simulation frames showed a stable trajectory. The average RMSD of the ligand S6U was 1.5 Å whereas, the protein chain and protein backbone also showed a stable RMSD throughout the 50 ns run with average RMSDs of 2.0 Å and 1.5 Å, respectively. Since the MD trajectory was stable at 50 ns further 100 ns run was not produced (Figure 4.13.).

(a)



(b)

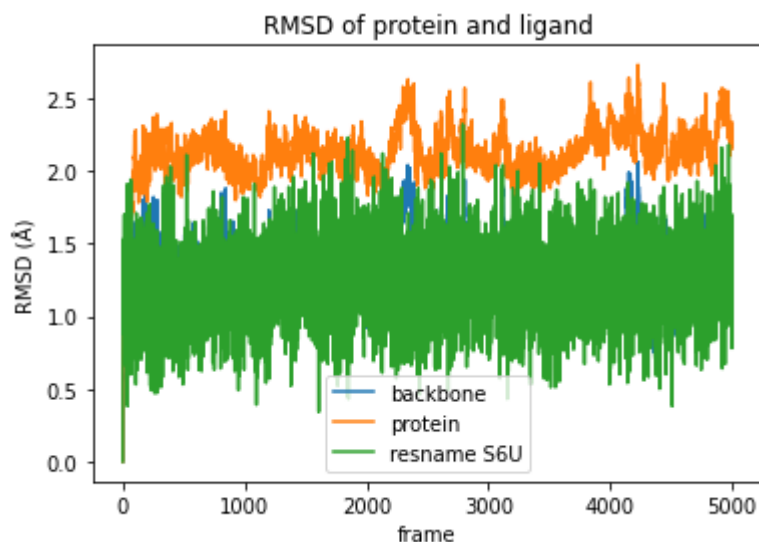
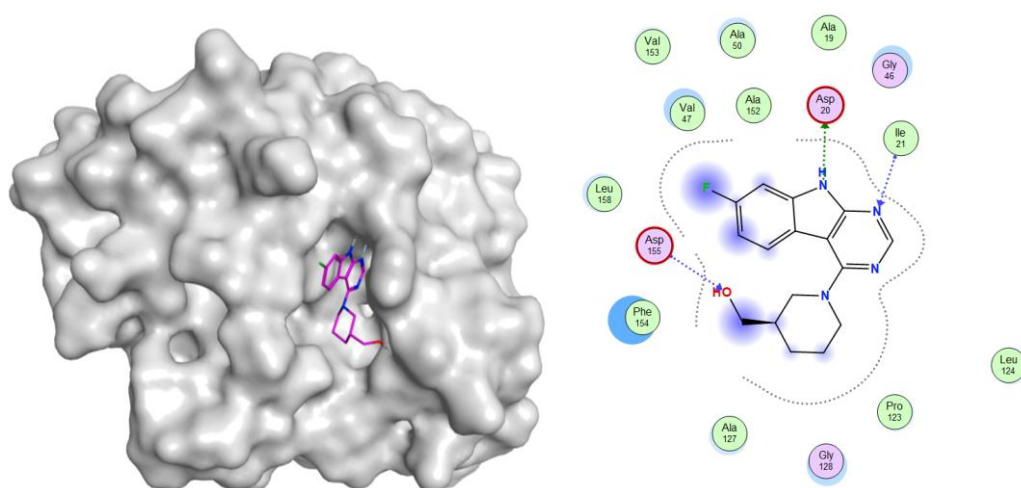


Figure 4.13. (a) Crystal structure of 5SPT with interaction map of S6U where the polar amino acid residues are highlighted in pink and the hydrophobic amino acid residues are highlighted in green. (b) RMSD plot of 5SPT during the 50 ns (5000 frames) MD simulation. The protein backbone fluctuations are shown in blue, the protein fluctuations are shown in orange and the ligand S6U fluctuations are shown in green. The x-axis represents MD simulation frames (5000 frames = 50 ns) and the y-axis represents RMSD fluctuation (Å).

The RMSD plot of the protein-RWQ complex (5SOJ) between the starting pose and the MD simulation frames showed a stable trajectory. The average RMSD of the ligand RWQ was 1.5 Å whereas, the protein chain and protein backbone also showed a stable RMSD throughout the 100 ns run with average RMSDs of 2.5 Å and 2.0 Å, respectively Figure 4.14.).

(a)



(b)

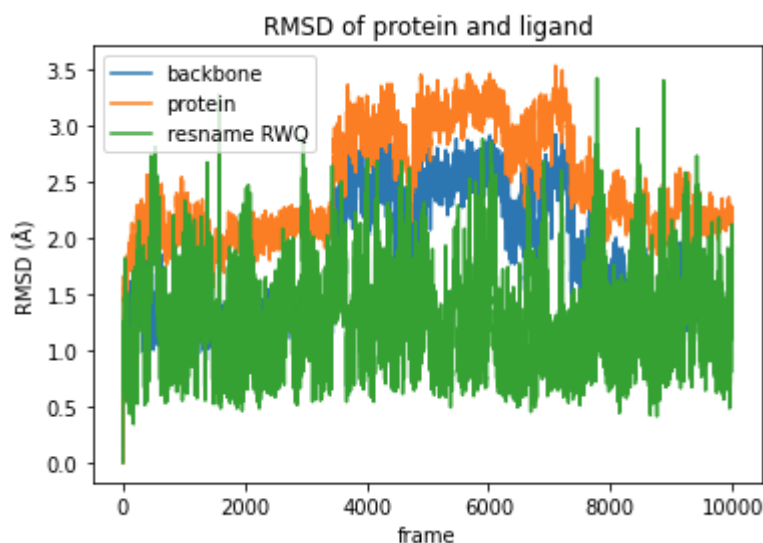


Figure 4.14. (a) Crystal structure of 5SOJ with interaction map RWQ where the polar amino acid residues are highlighted in pink and the hydrophobic amino acid residues are highlighted in green. (b) RMSD plot of 5SOJ during the 100 ns (10000 frames) MD simulation. The protein backbone fluctuations are shown in blue, the protein fluctuations are shown in orange and the ligand RWQ fluctuations are shown in green. The x-axis represents MD simulation frames (10000 frames = 100 ns) and the y-axis represents RMSD fluctuation (Å).

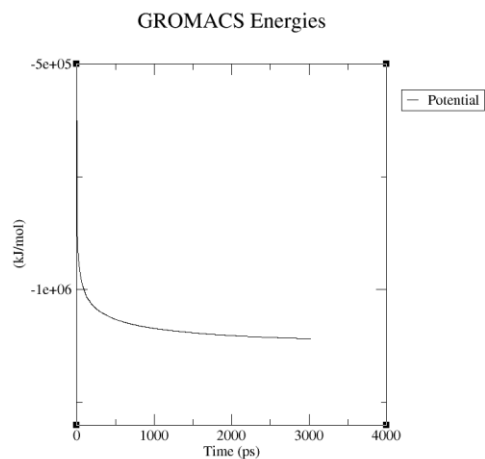
4.8.3. MD on CACHE SC-2 Mac1-NSP3 hits

The FRED docked poses of the two hits, CACHE3HI_1689_15 ($K_d = 328 \mu\text{M}$), referred to as SK1, and CACHE3HI_1689_35 ($K_d = 4.8 \mu\text{M}$), referred to as SK2, were used to carry out the MD simulations. The ligand parameterisation with CGenFF resulted in a parameter penalty of 29 and a charge penalty of 14.9 for the ligand SK1. SK2 had parameter and charge penalties of 24.0 and 37.06, respectively. The files obtained during the MD simulations were inspected. System stability was checked for the two protein-ligand complexes by analysing the system's potential energy, average temperature, average pressure, RMSF, density and Rg. To analyse the trajectories, an RMSD graph was plotted between the starting structures and the MD frames followed by hydrogen bond analysis and structural overlays of the MD representative frames with the docked poses.

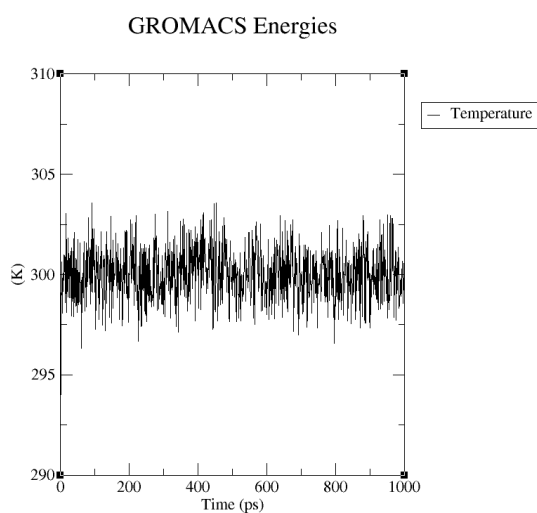
For the **protein-SK1 complex**, the average potential energy after energy minimisation was $-1.08\text{e}+06$, the average temperature of the system was maintained at 299.97 K, the average pressure of the system was maintained at 0.66 bar, no major RMSF fluctuations were

observed of the amino acid residues, the density of the system was maintained at 1017.64 kg/m³, and the Rg plot showed compactness of the complex with structural stability and slight structural fluctuations (Figure 4.15.).

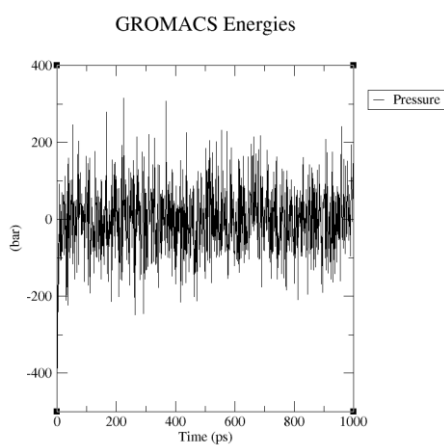
(a)



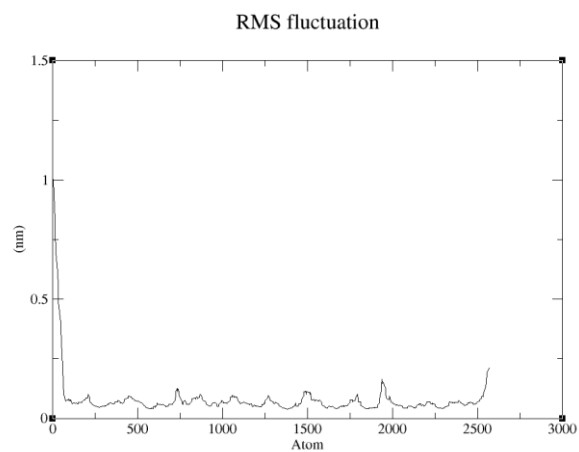
(b)



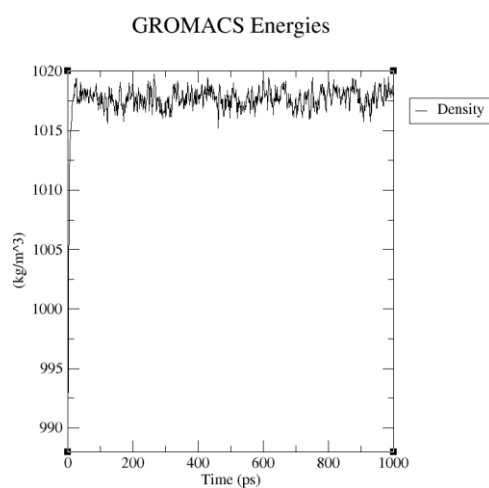
(c)



(d)



(e)



(f)

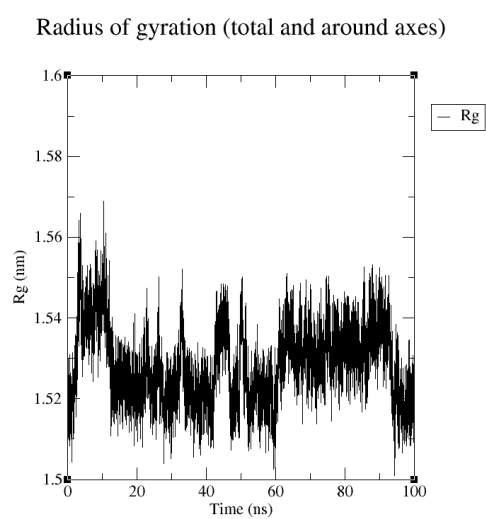
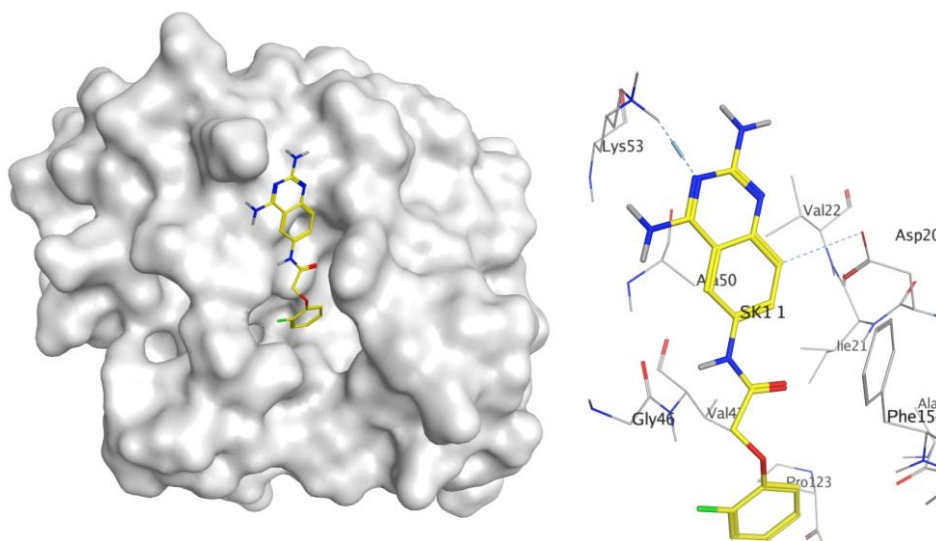


Figure 4.15. (a) Protein-SK1 complex average potential energy plot of the system. The x-axis shows time (ps) and y-axis shows potential energy (kJ/mol). (b) Protein-SK1 complex

average temperature plot of the system. The x-axis shows time (ps) and y-axis shows temperature (K). (c) Protein-SK1 complex average pressure plot of the system. The x-axis shows time (ps) and y-axis shows pressure (bar). (d) RMSF plot of each residue atom of the SC-2 Mac1-NSP3 protein. The x-axis shows atom position and y-axis shows atom fluctuations (nm). (e) Protein-SK1 complex average density plot of the system. The x-axis shows time (ps) and y-axis shows density (kg/m^3). (f) Rg plot of the protein-SK1 complex. The x-axis shows time (ns) and y-axis shows fluctuations (nm).

The FRED docked pose of SK1 with SC-2 Mac1-NSP3 protein showed that there were hydrogen bond interactions with ASP20 and LYS53, which were in the intermediate conserved amino acid residue threshold as per the analysis done in Chapter III (Figure 4.16.). The xyz co-ordinates of the docked pose were the input to the MD simulation and were also the first frame in the MD trajectory.

(a)



(b)

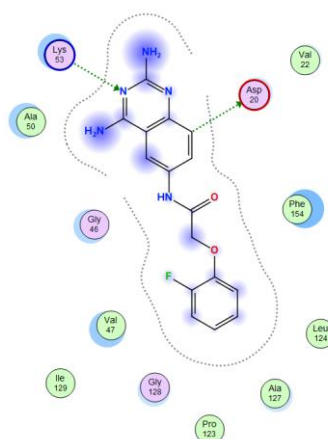
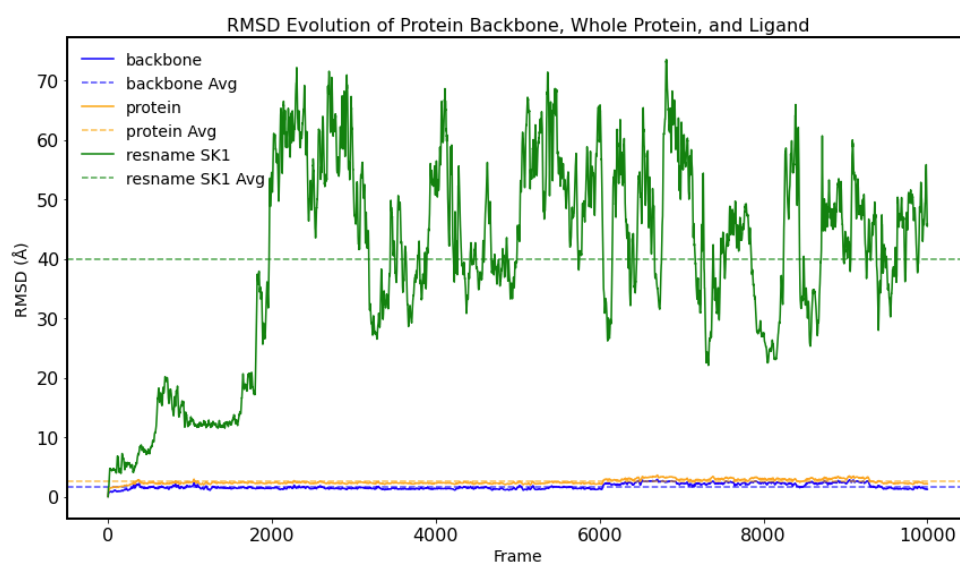


Figure 4.16. (a) FRED docked pose of SK1 with hydrogen bond interactions (dotted lines) with LYS53 and ASP20 of SC-2 Mac1-NSP3. (b) Interaction map of SK1 where the polar amino acid residues are highlighted in pink and the hydrophobic amino acid residues are highlighted in green.

The RMSD plot of the protein-SK1 complex between the starting pose and the MD simulation frames showed a high level of fluctuations, which indicates unbinding of the ligand. The protein chain and protein backbone showed a stable RMSD throughout the 100 ns run, but the ligand (SK1) left the binding pocket at approximately 20 ns (Figure 4.17.). Since the MD trajectory showed unbinding of SK1, it was not selected for further ABFE analysis.

(a)



(b)

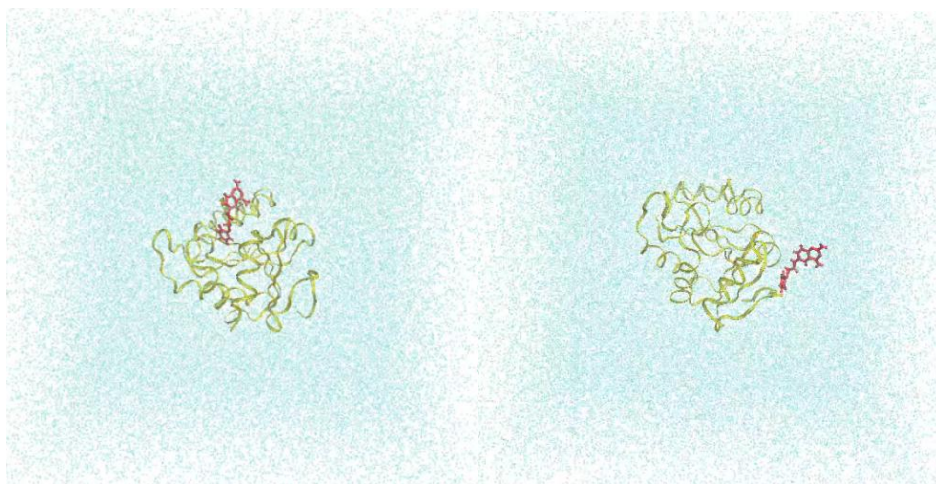


Figure 4.17. (a) RMSD plot of protein-SK1 during the 100 ns (10000 frames) MD

simulation. The protein backbone fluctuations are shown in blue, the protein fluctuations are shown in yellow and the ligand SK1 fluctuations are shown in green. The average fluctuations for the three are shown in dotted lines. The x-axis represents MD simulation frames (10000 frames = 100 ns) and the y-axis represents RMSD fluctuation (Å). (b) Protein-SK1 complex in cubic solvent box showing the initial pose and the frame where SK1 leaves the binding pocket. Protein is represented in yellow, SK1 in red and water molecules in blue.

The hydrogen bond analysis on the protein-SK1 complex revealed that hydrogen bonds with the LYS53 and ASP20 residues were not maintained during the 100 ns MD simulation, but instead there were new hydrogen bonds formed due to conformational changes of the ligand. The new hydrogen bonds formed were with PHE154, which was maintained for 4.75% of the MD run, and VAL47, which was maintained for 1.44% of the MD run (Figure 4.18.).

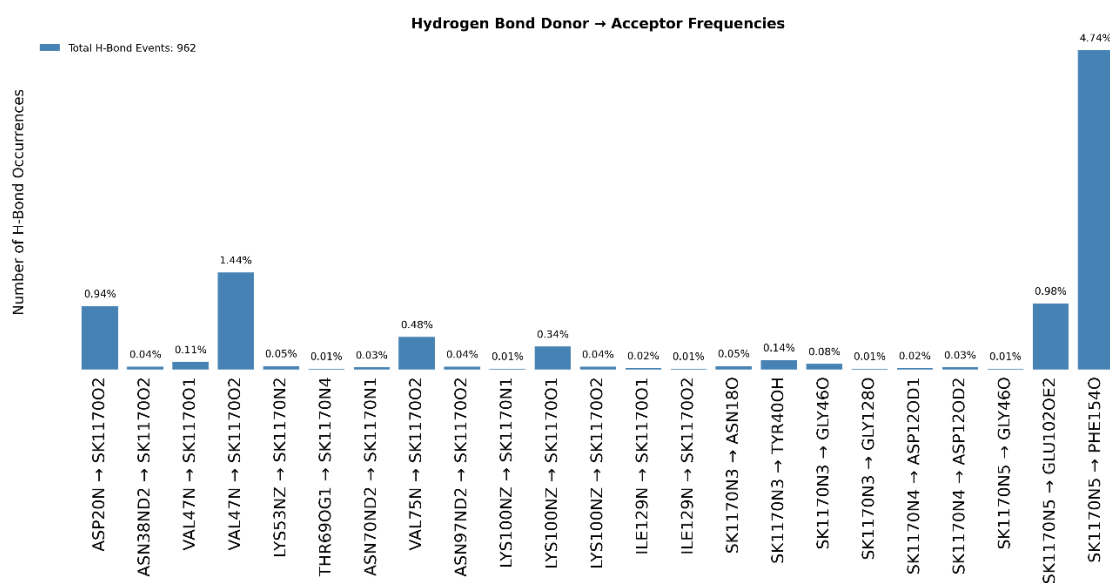


Figure 4.18. Hydrogen bond analysis showing ligand SK1 and protein atoms involved in the formation of hydrogen bonds. The x-axis shows the SK1, and protein atoms involved in hydrogen bond formation and y-axis shows number of H-bonds (%).

Lastly, the structural overlay between the MD representative frame and the docked pose showed an RMSD of 2.1 Å between the proteins and an RMSD of 24.7 Å between the ligand poses (Figure 4.19.). The high RMSD between the ligand poses reflected the unbinding of SK1 from the protein binding pocket.

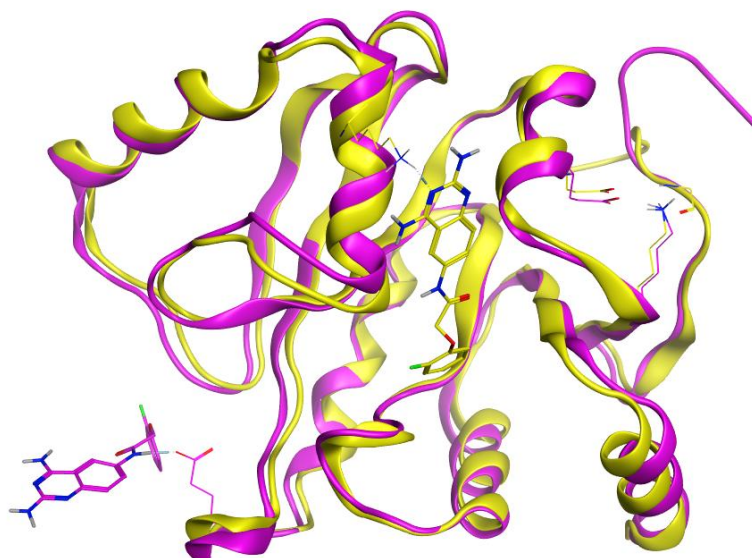
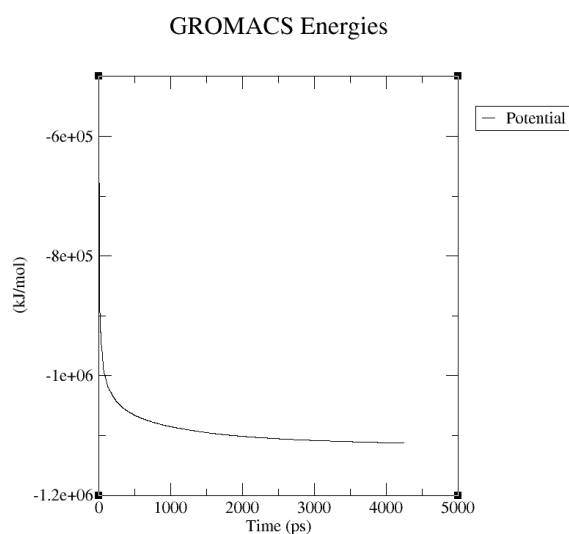


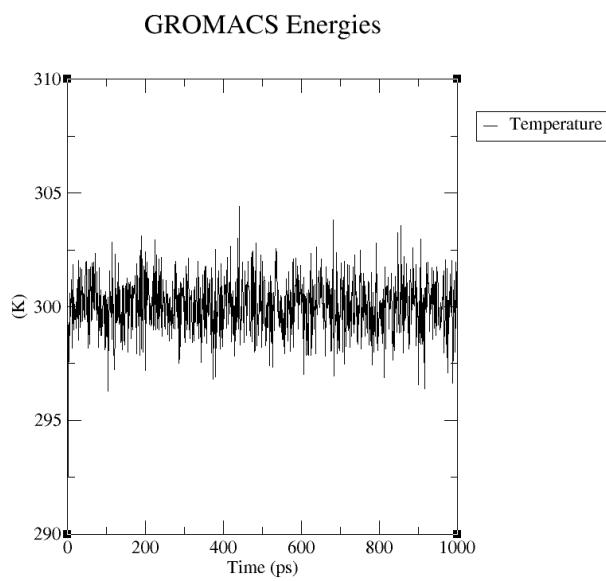
Figure 4.19. Structural overlay between the MD representative frame (pink) and the docked pose (yellow) of protein-SK1 complex.

For the **protein-SK2 complex**, the average potential energy after energy minimisation was -1.09×10^6 , the average temperature of the system was maintained at 299.98 K, the average pressure of the system was maintained at 0.74 bar, no major RMSF fluctuations were observed of the amino acid residues, the density of the system was maintained at 1017.61 kg/m^3 , and the Rg plot showed compactness of the complex with structural stability and slight structural fluctuations (Figure 4.20.).

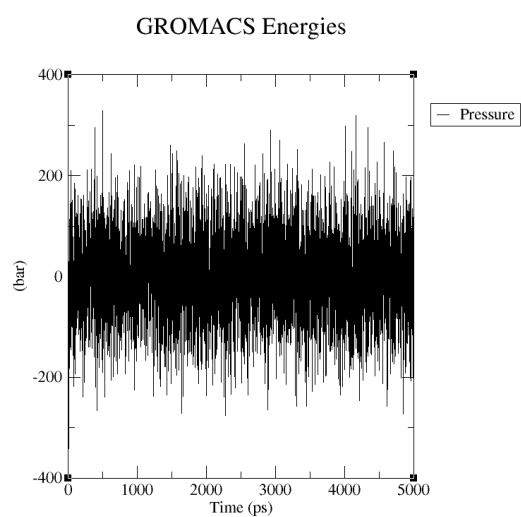
(a)



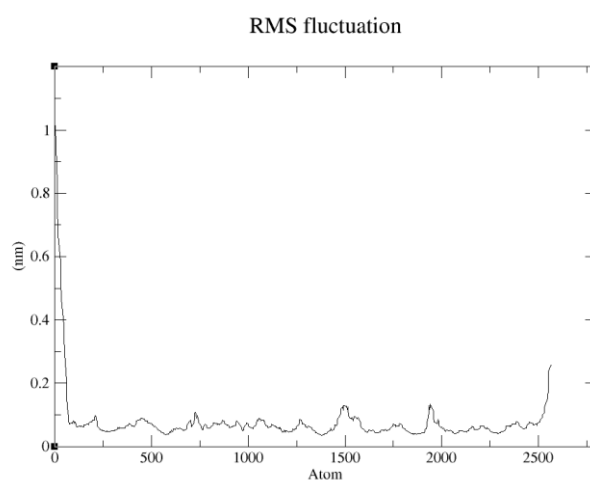
(b)



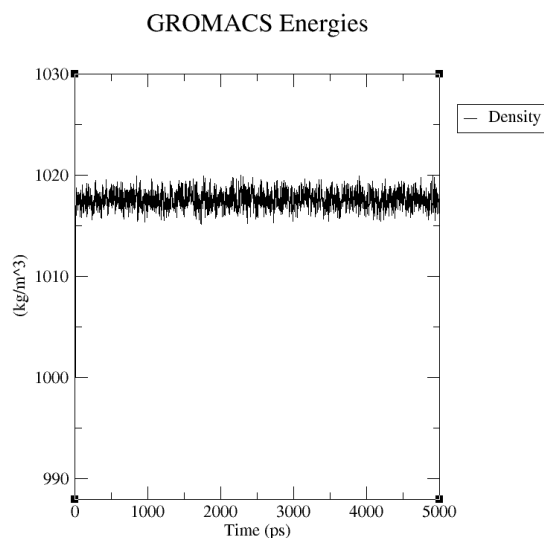
(c)



(d)



(e)



(f)

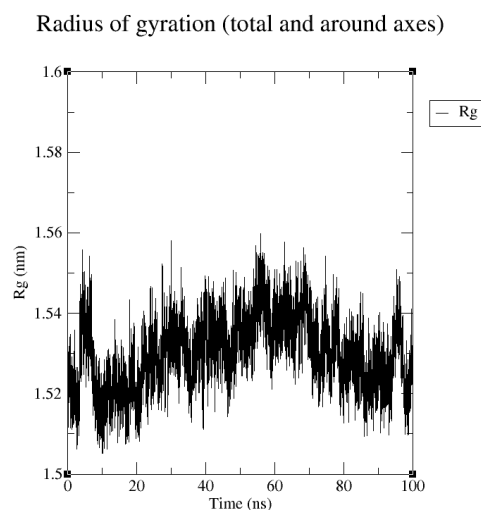


Figure 4.20. (a) Protein-SK2 complex average potential energy plot of the system. The x-axis shows time (ps) and y-axis shows potential energy (kJ/mol). (b) Protein-SK2 complex average temperature plot of the system. The x-axis shows time (ps) and y-axis shows temperature (K). (c) Protein-SK2 complex average pressure plot of the system. The x-axis shows time (ps) and y-axis shows pressure (bar). (d) RMSF plot of each residue atom of the SC-2 Mac1-NSP3 protein. The x-axis shows atom position and y-axis shows atom fluctuations (nm). (e) Protein-SK2 complex average density plot of the system. The x-axis shows time (ps) and y-axis shows density (kg/m³). (f) Rg plot of the protein-SK2 complex. The x-axis shows time (ns) and y-axis shows fluctuations (nm).

The FRED docked pose of SK2 with SC-2 Mac1-NSP3 protein showed that there were hydrogen bond interactions with ASP20, ILE121 and LYS53, which were in the conserved

and intermediate conserved amino acid residue thresholds as per the analysis done in Chapter III (Figure 4.21.). The xyz co-ordinates of the docked pose were the input to the MD simulation and were also the first frame in the MD trajectory.

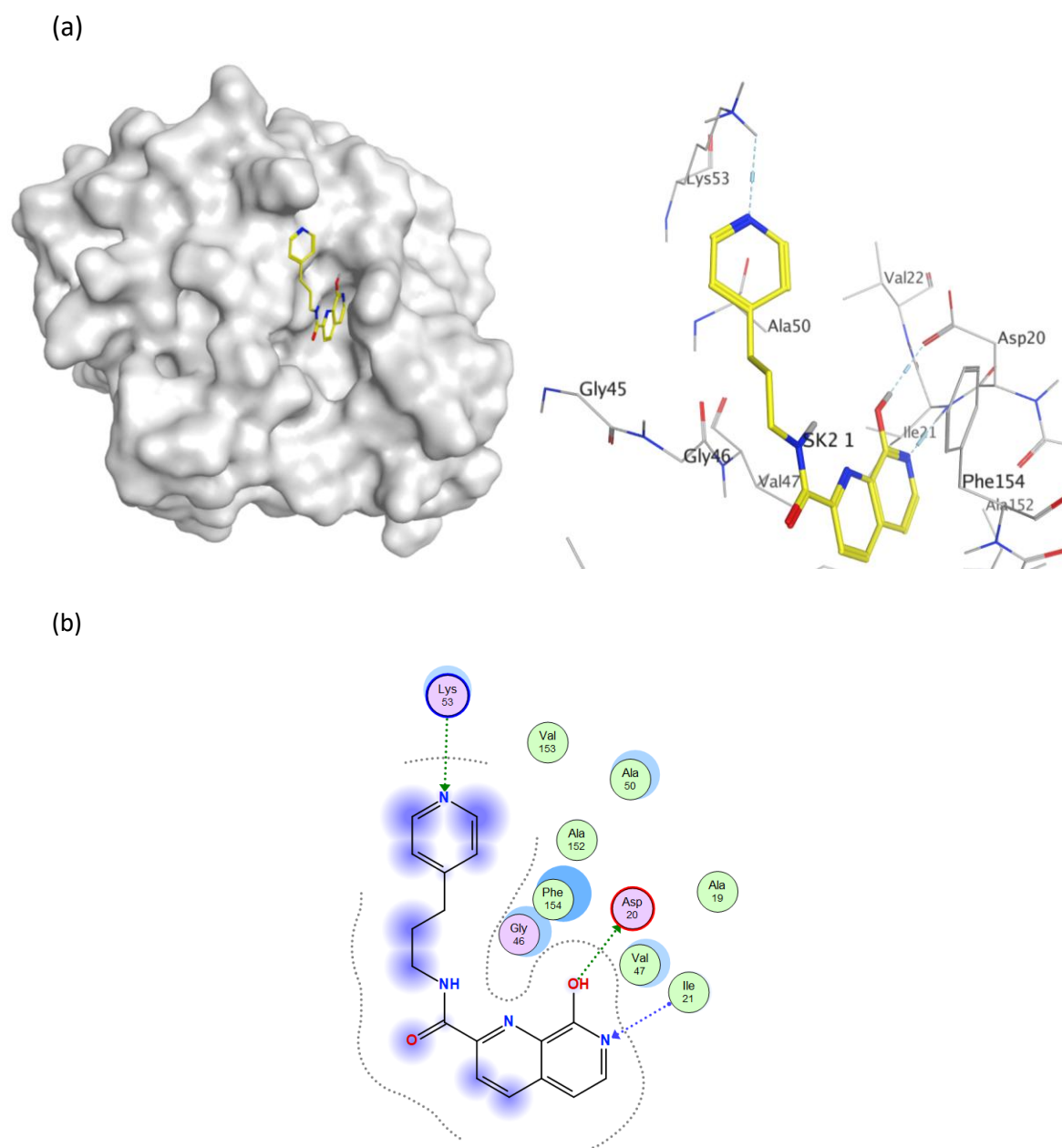
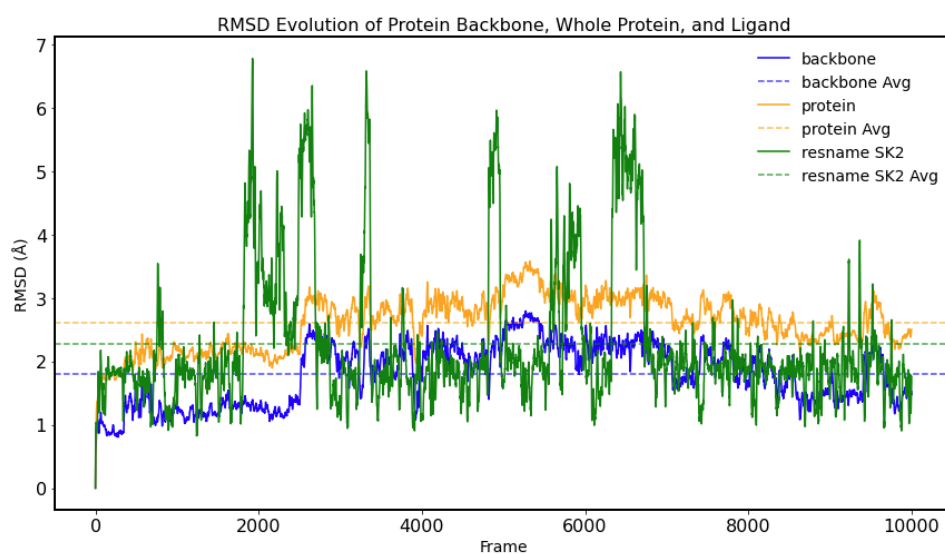


Figure 4.21. (a) FRED docked pose of SK2 with hydrogen bond interactions (dotted lines) with ASP20, ILE21 and LYS53 of SC-2 Mac1-NSP3. (b) Interaction map of SK2 where the polar amino acid residues are highlighted in pink and the hydrophobic amino acid residues are highlighted in green.

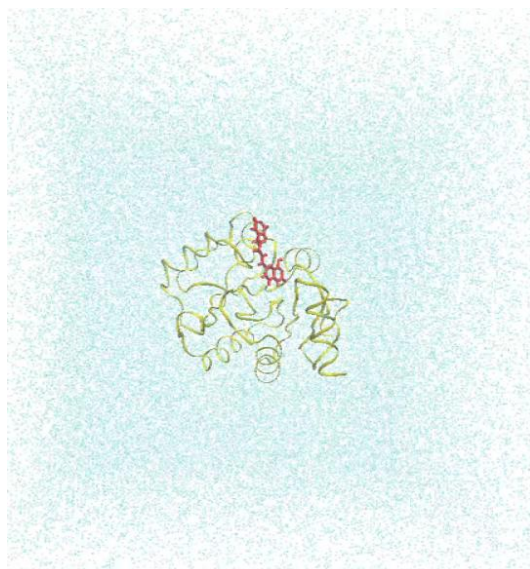
The RMSD plot of the protein-SK2 complex between the starting pose and the MD simulation frames showed a stable trajectory. The average RMSD of the ligand SK2 was 2.5 Å with a few fluctuations, which indicates conformational changes. The protein chain and

protein backbone also showed a stable RMSD throughout the 100 ns run with average RMSDs of 2.8 Å and 1.9 Å, respectively. To further explore the conformational change of the ligand SK2, an MD run of 500 ns also revealed that the ligand stays within the binding pocket but undergoes continuous conformational changes with an average RMSD of 3.8 Å (Figure 4.22.).

(a)



(b)



(c)

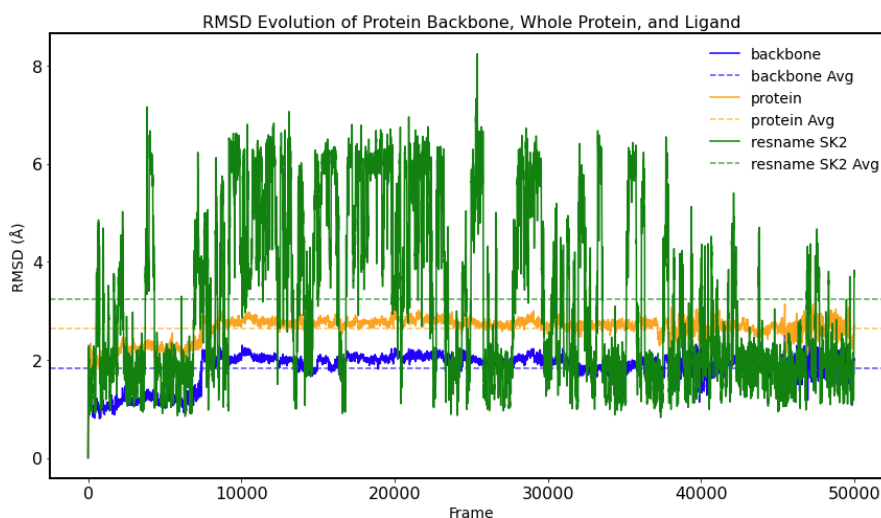


Figure 4.22. (a) RMSD plot of protein-SK2 during the 100 ns (10000 frames) MD simulation. The protein backbone fluctuations are shown in blue, the protein fluctuations are shown in yellow and the ligand SK2 fluctuations are shown in green. The average fluctuations for the three are shown in dotted lines. The x-axis represents MD simulation frames (10000 frames = 100 ns) and the y-axis represents RMSD fluctuation (Å). (b) Protein-SK2 complex in cubic solvent box showing the initial pose. Protein is represented in yellow, SK1 in red and water molecules in blue. (c) RMSD plot of protein-SK2 during the 500 ns (50000 frames) MD simulation. The protein backbone fluctuations are shown in blue, the protein fluctuations are shown in yellow and the ligand SK2 fluctuations are shown in green. The average fluctuations for the three are shown in dotted lines. The x-axis represents MD simulation frames (50000 frames = 500 ns) and the y-axis represents RMSD fluctuation (Å).

The hydrogen bond analysis on the protein-SK2 complex revealed that the hydrogen bond with the LYS53 was not maintained during the 100 ns MD simulation. This is the reason why the ligand SK2 shows spikes in the RMSD plots, the absence of the hydrogen bond allows the ligand to acquire different conformation. On the other hand, the hydrogen bonds formed with ASP20 were maintained for 87.86% of the MD run, and with ILE21, they were maintained for 93.09% of the MD run (Figure 4.23.).

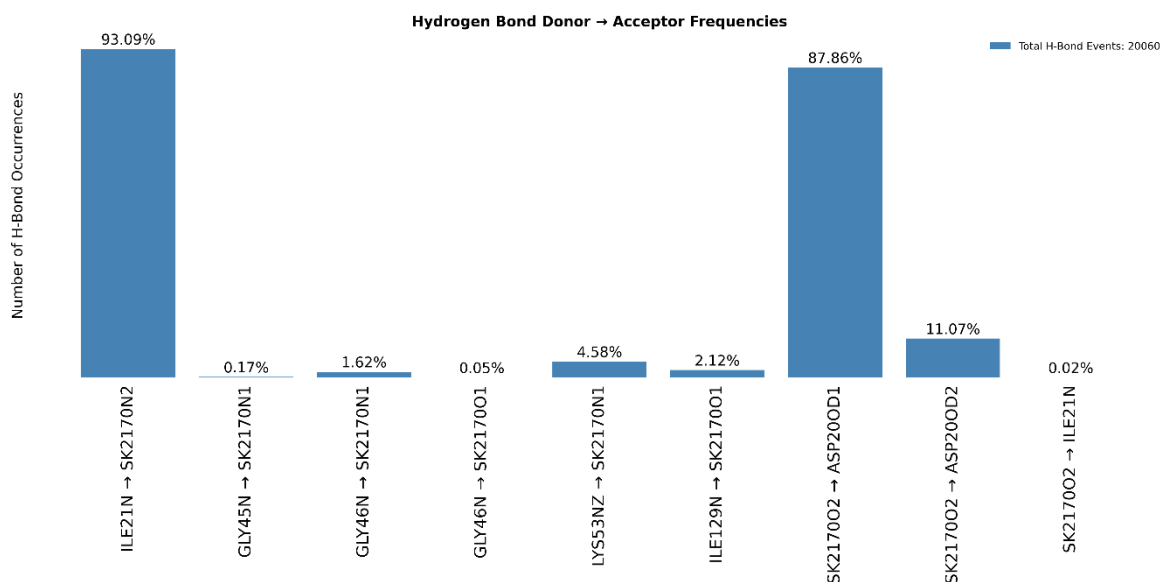


Figure 4.23. Hydrogen bond analysis showing ligand SK2 and protein atoms involved in the formation of hydrogen bonds. The x-axis shows the SK2, and protein atoms involved in hydrogen bond formation and y-axis shows number of H-bonds (%).

Lastly, the structural overlay between the MD representative frame and the docked pose showed an RMSD of 2.1 Å between the proteins and an RMSD of 0.97 Å between the ligand poses (Figure 4.24.). This showed that both MD and docking predicted the same binding mode of SK2 with SC-2 Mac1-NSP3 protein.

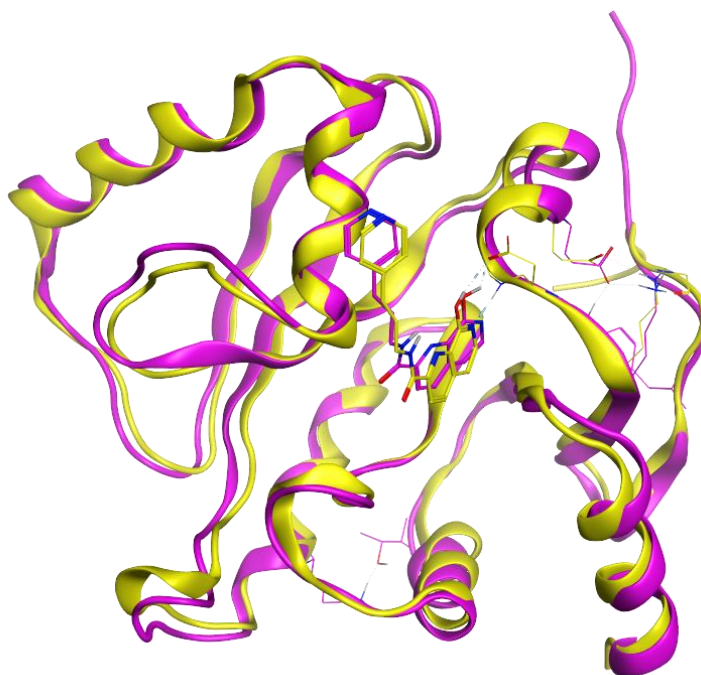


Figure 4.24. Structural overlay between the MD representative frame (pink) and the docked pose (yellow) of protein-SK2 complex.

4.8.4. Benchmarking results of ABFE protocol on SC-2 M_{pro} (NSP5)

The ABFE protocol by Ghahremanpour et al., 2023, was replicated on the SC-2 M_{pro} (NSP5) protein. The input files provided (PDB: 5R82) (Figure 4.25. a) were used to carry out the bound and unbound state simulations. The GROMACS bar module was used to calculate the free energies for each state by extracting the md.xvg files from each λ window run. The energy values were converted from kJ/mol to kcal/mol. The free energies of the bound state and unbound states were 28.6 kcal/mol and 14.03 kcal/mol, respectively (Figure 4.25 b, c). A standard-state correction of approximately 6.5 kcal/mol was added to account for the Boresch restraints. After computing these values to the following equation, the ΔG_{pred}^0 binding free energy value was -7.59 kcal/mol.

$$\Delta G_{pred}^0 = -28.17 + 14.03 + 6.5$$

$$\Delta G_{pred}^0 = -7.59 \text{ kcal/mol}$$

The predicted ΔG_{pred}^0 slightly overestimates the binding free energy compared to the experimental ΔG_{exp}^0 i.e., -6.05 kcal/mol. The workflow by Ghahremanpour et al., 2023 had also predicted the same value of ΔG_{pred}^0 i.e., -7.59 kcal/mol. Thus, the replication of the workflow helped in validating the capacity of GROMACS to carry out an ABFE calculation.

(a)



(b)

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 12.22 +/- 0.17	point	0 -	1,	DG 7.48 +/- 0.06
point	1 -	2,	DG 11.71 +/- 0.05	point	1 -	2,	DG 7.22 +/- 0.08
point	2 -	3,	DG 10.53 +/- 0.06	point	2 -	3,	DG 6.83 +/- 0.13
point	3 -	4,	DG 9.07 +/- 0.12	point	3 -	4,	DG 6.65 +/- 0.15
point	4 -	5,	DG 7.99 +/- 0.20	point	4 -	5,	DG 6.69 +/- 0.06
point	5 -	6,	DG 6.45 +/- 0.11	point	5 -	6,	DG 6.45 +/- 0.10
point	6 -	7,	DG 5.27 +/- 0.17	point	6 -	7,	DG 5.92 +/- 0.22
point	7 -	8,	DG 4.15 +/- 0.07	point	7 -	8,	DG 5.47 +/- 0.25
point	8 -	9,	DG 2.62 +/- 0.17	point	8 -	9,	DG 5.70 +/- 0.19
point	9 -	10,	DG 1.96 +/- 0.11	point	9 -	10,	DG 5.08 +/- 0.12
total	0 -	10,	DG 71.96 +/- 0.46	point	10 -	11,	DG 3.77 +/- 0.20
Bound_coul = 17.18 kcal/mol				point	11 -	12,	DG 2.90 +/- 0.40
Complex= 28.6 kcal/mol				point	12 -	13,	DG 1.61 +/- 0.14
(vdw+coul+restr)				point	13 -	14,	DG -0.63 +/- 0.18
				point	14 -	15,	DG -4.11 +/- 0.33
				point	15 -	16,	DG -9.23 +/- 0.12
				point	16 -	17,	DG -9.67 +/- 0.12
				point	17 -	18,	DG -6.67 +/- 0.16
				point	18 -	19,	DG -3.56 +/- 0.04
				point	19 -	20,	DG -0.14 +/- 0.07
				total	0 -	20,	DG 37.77 +/- 1.31
				Bound_restr = 2.4 kcal/mol			
				Bound_vdw = 9.02 kcal/mol			

(c)

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 13.66 +/- 0.04	point	0 -	1,	DG 4.82 +/- 0.01
point	1 -	2,	DG 11.82 +/- 0.03	point	1 -	2,	DG 4.72 +/- 0.01
point	2 -	3,	DG 9.99 +/- 0.05	point	2 -	3,	DG 4.60 +/- 0.01
point	3 -	4,	DG 8.15 +/- 0.02	point	3 -	4,	DG 4.47 +/- 0.01
point	4 -	5,	DG 6.52 +/- 0.03	point	4 -	5,	DG 4.32 +/- 0.01
point	5 -	6,	DG 5.11 +/- 0.03	point	5 -	6,	DG 4.14 +/- 0.01
point	6 -	7,	DG 3.79 +/- 0.01	point	6 -	7,	DG 3.92 +/- 0.01
point	7 -	8,	DG 2.56 +/- 0.01	point	7 -	8,	DG 3.66 +/- 0.00
point	8 -	9,	DG 1.40 +/- 0.01	point	8 -	9,	DG 3.32 +/- 0.00
point	9 -	10,	DG 0.27 +/- 0.01	point	9 -	10,	DG 2.84 +/- 0.01
total	0 -	10,	DG 63.28 +/- 0.08	point	10 -	11,	DG 2.16 +/- 0.02
Unbound_coul = 15.12 kcal/mol				point	11 -	12,	DG 1.21 +/- 0.03
Ligand in solvent = 14.03 kcal/mol				point	12 -	13,	DG -0.25 +/- 0.01
(vdw+coul)				point	13 -	14,	DG -2.65 +/- 0.05
				point	14 -	15,	DG -6.77 +/- 0.08
				point	15 -	16,	DG -11.50 +/- 0.08
				point	16 -	17,	DG -11.90 +/- 0.04
				point	17 -	18,	DG -8.90 +/- 0.02
				point	18 -	19,	DG -5.14 +/- 0.02
				point	19 -	20,	DG -1.62 +/- 0.01
				total	0 -	20,	DG -4.57 +/- 0.18
				Unbound_vdw = -1.09 kcal/mol			

Figure 4.25. (a) Crystal structure of SC-2 M_{pro} (blue) (PDB: 5R82) in complex with Z219104216 (red). (b) Snapshot of the bound state free energy differences across the various windows of electrostatic (11 windows) and vdw decoupling (21 windows) along with restraints contribution (12 windows). (c) Snapshot of the unbound state free energy differences across the various windows of electrostatic (11 windows) and vdw decoupling (21 windows). The free energy difference (DG- delta G) between two consecutive windows are calculated in kJ/mol. The cumulative free energy (kcal/mol) is calculated by adding the contribution of each point (2 consecutive windows) for each decoupling state.

4.8.5. ABFE results on SC-2 Mac1-NSP3

To validate the ABFE protocol, the known binders S6U and RWQ from the **PDBs 5SPT and 5SOJ** were used as positive controls when evaluating **CACHE3HI_1689_35**. The K_i and K_d values were used for calculating the experimental binding free energies. The md.gro files from the complex in solvent and ligand in solvent MD runs were used as the input files to ABFE calculations since the systems were energy minimised and well equilibrated. For the three protein-ligand complexes, the ABFE calculations were run in triplicate.

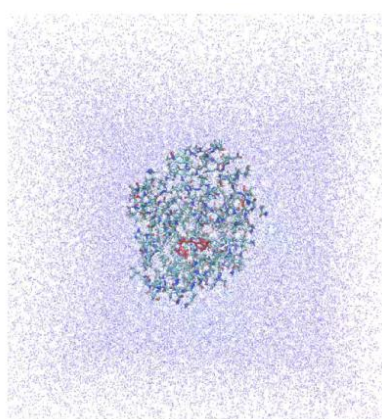
For the first round of ABFE run on **5SPT** (Figure 4.26. a), the free energies of the bound state and unbound states were 34.89 kcal/mol and 21.04 kcal/mol, respectively (Figure 4.26. b, c). A standard-state correction of approximately 7.01 kcal/mol was added to account for the Boresch restraints which was calculated by the MDRestrainsGenerator. After computing these values to the following equation, the ΔG_{pred}^0 binding free energy value was -6.84 kcal/mol.

$$\Delta G_{pred}^0 = -34.89 + 21.04 + 7.01$$

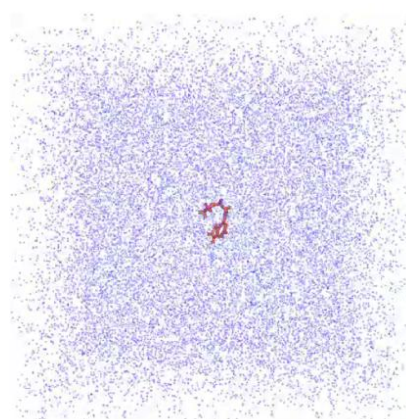
$$\Delta G_{pred}^0 = -6.84 \text{ kcal/mol}$$

The predicted ΔG_{pred}^0 slightly overestimates the binding free energy compared to the experimental ΔG_{exp}^0 i.e., -5.84 kcal/mol. The ΔG_{pred}^0 binding free energy values of second and the third runs were -6.80 kcal/mol and -6.79 kcal/mol, respectively (**Appendix 6.6.**). The coulombic decoupling in the bound state showed a smooth, gradual trend in the free energy differences across the λ windows, while the van der Waals decoupling followed a bell-shaped curve (Figure 4.26. d).

(a)



Complex (protein+ligand) in solvent



Ligand in solvent

(b)

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 23.34 +/- 0.13	point	0 -	1,	DG 4.55 +/- 0.05
point	1 -	2,	DG 21.26 +/- 0.12	point	1 -	2,	DG 4.39 +/- 0.06
point	2 -	3,	DG 19.16 +/- 0.14	point	2 -	3,	DG 4.14 +/- 0.06
point	3 -	4,	DG 17.48 +/- 0.07	point	3 -	4,	DG 4.02 +/- 0.07
point	4 -	5,	DG 14.47 +/- 0.23	point	4 -	5,	DG 4.04 +/- 0.09
point	5 -	6,	DG 10.72 +/- 0.37	point	5 -	6,	DG 3.75 +/- 0.08
point	6 -	7,	DG 8.80 +/- 0.35	point	6 -	7,	DG 3.55 +/- 0.08
point	7 -	8,	DG 7.23 +/- 0.06	point	7 -	8,	DG 3.53 +/- 0.09
point	8 -	9,	DG 5.20 +/- 0.13	point	8 -	9,	DG 3.37 +/- 0.11
point	9 -	10,	DG 3.73 +/- 0.19	point	9 -	10,	DG 2.64 +/- 0.14
total	0 -	10,	DG 131.39 +/- 1.26	point	10 -	11,	DG 2.06 +/- 0.16
Bound_coul = 31.40 kcal/mol				point	11 -	12,	DG 1.77 +/- 0.08
Complex = 34.89 kcal/mol				point	12 -	13,	DG 0.66 +/- 0.18
				point	13 -	14,	DG -1.21 +/- 0.20
				point	14 -	15,	DG -4.35 +/- 0.52
				point	15 -	16,	DG -7.85 +/- 0.36
				point	16 -	17,	DG -7.89 +/- 0.14
				point	17 -	18,	DG -5.66 +/- 0.10
				point	18 -	19,	DG -2.97 +/- 0.08
				point	19 -	20,	DG -0.02 +/- 0.13
				total	0 -	20,	DG 12.54 +/- 1.38
				Final results in kJ/mol:			
				point	0 -	1,	DG 0.03 +/- 0.00
				point	1 -	2,	DG 0.04 +/- 0.00
				point	2 -	3,	DG 0.06 +/- 0.00
				point	3 -	4,	DG 0.07 +/- 0.00
				point	4 -	5,	DG 0.06 +/- 0.00
				point	5 -	6,	DG 0.12 +/- 0.00
				point	6 -	7,	DG 0.12 +/- 0.00
				point	7 -	8,	DG 0.34 +/- 0.00
				point	8 -	9,	DG 0.32 +/- 0.01
				point	9 -	10,	DG 0.52 +/- 0.02
				point	10 -	11,	DG 0.46 +/- 0.01
				total	0 -	11,	DG 2.13 +/- 0.05
				Bound_restr = 0.50 kcal/mol			
				Bound_vdw = 2.99 kcal/mol			

(c)

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 28.75 +/- 0.14	point	0 -	1,	DG 3.40 +/- 0.01
point	1 -	2,	DG 17.44 +/- 0.15	point	1 -	2,	DG 3.35 +/- 0.01
point	2 -	3,	DG 14.43 +/- 0.07	point	2 -	3,	DG 3.27 +/- 0.01
point	3 -	4,	DG 11.83 +/- 0.07	point	3 -	4,	DG 3.17 +/- 0.01
point	4 -	5,	DG 9.43 +/- 0.05	point	4 -	5,	DG 3.04 +/- 0.01
point	5 -	6,	DG 7.31 +/- 0.03	point	5 -	6,	DG 2.90 +/- 0.01
point	6 -	7,	DG 5.47 +/- 0.02	point	6 -	7,	DG 2.72 +/- 0.01
point	7 -	8,	DG 3.75 +/- 0.02	point	7 -	8,	DG 2.50 +/- 0.01
point	8 -	9,	DG 2.17 +/- 0.01	point	8 -	9,	DG 2.22 +/- 0.01
point	9 -	10,	DG 0.69 +/- 0.01	point	9 -	10,	DG 1.81 +/- 0.01
total	0 -	10,	DG 93.25 +/- 0.49	point	10 -	11,	DG 1.31 +/- 0.01
Unbound_coul = 22.28 kcal/mol				point	11 -	12,	DG 0.66 +/- 0.01
Ligand in solvent = 21.04 kcal/mol				point	12 -	13,	DG -0.29 +/- 0.03
				point	13 -	14,	DG -1.68 +/- 0.03
				point	14 -	15,	DG -4.01 +/- 0.06
				point	15 -	16,	DG -7.45 +/- 0.05
				point	16 -	17,	DG -8.99 +/- 0.03
				point	17 -	18,	DG -7.33 +/- 0.02
				point	18 -	19,	DG -4.39 +/- 0.01
				point	19 -	20,	DG -1.41 +/- 0.01
				total	0 -	20,	DG -5.19 +/- 0.14
				Unbound_vdw = -1.24 kcal/mol			

(d)

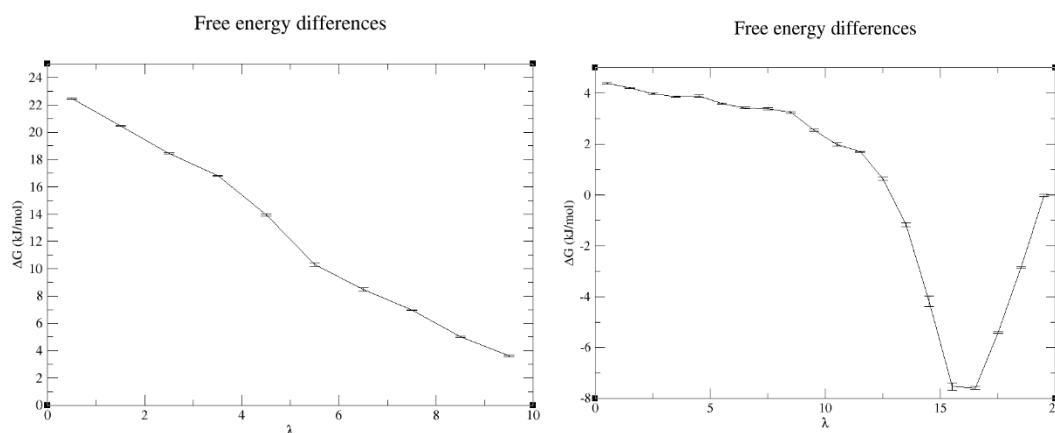


Figure 4.26. (a) Bound and unbound simulation box showing complex (5SPT) in solvent and ligand in solvent (S6U) respectively. The ligand is shown in red; the protein is shown

(b)

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 22.56 +/- 0.03	point	0 -	1,	DG 4.04 +/- 0.00
point	1 -	2,	DG 19.40 +/- 0.01	point	1 -	2,	DG 3.96 +/- 0.01
point	2 -	3,	DG 16.37 +/- 0.02	point	2 -	3,	DG 3.88 +/- 0.01
point	3 -	4,	DG 13.52 +/- 0.02	point	3 -	4,	DG 3.77 +/- 0.01
point	4 -	5,	DG 10.81 +/- 0.02	point	4 -	5,	DG 3.64 +/- 0.00
point	5 -	6,	DG 8.33 +/- 0.01	point	5 -	6,	DG 3.47 +/- 0.01
point	6 -	7,	DG 6.12 +/- 0.01	point	6 -	7,	DG 3.28 +/- 0.01
point	7 -	8,	DG 4.18 +/- 0.02	point	7 -	8,	DG 3.05 +/- 0.01
point	8 -	9,	DG 2.41 +/- 0.02	point	8 -	9,	DG 2.74 +/- 0.01
point	9 -	10,	DG 0.75 +/- 0.01	point	9 -	10,	DG 2.31 +/- 0.02
total	0 -	10,	DG 104.46 +/- 0.07	point	10 -	11,	DG 1.78 +/- 0.01
Unbound_coul = 24.96 kcal/mol				point	11 -	12,	DG 1.02 +/- 0.01
				point	12 -	13,	DG 0.01 +/- 0.03
				point	13 -	14,	DG -1.51 +/- 0.03
				point	14 -	15,	DG -4.36 +/- 0.03
				point	15 -	16,	DG -8.82 +/- 0.05
				point	16 -	17,	DG -10.62 +/- 0.06
				point	17 -	18,	DG -8.51 +/- 0.02
				point	18 -	19,	DG -5.09 +/- 0.01
				point	19 -	20,	DG -1.63 +/- 0.01
				total	0 -	20,	DG -3.59 +/- 0.13
Ligand in solvent= 24.11 kcal/mol				Unbound_vdw = -0.85 kcal/mol			

Figure 4.27. (a) Snapshot of the bound state free energy differences across the various windows of electrostatic and vdw decoupling along with restraints contribution. (b) Snapshot of the unbound state free energy differences across the various windows of electrostatic and vdw decoupling.

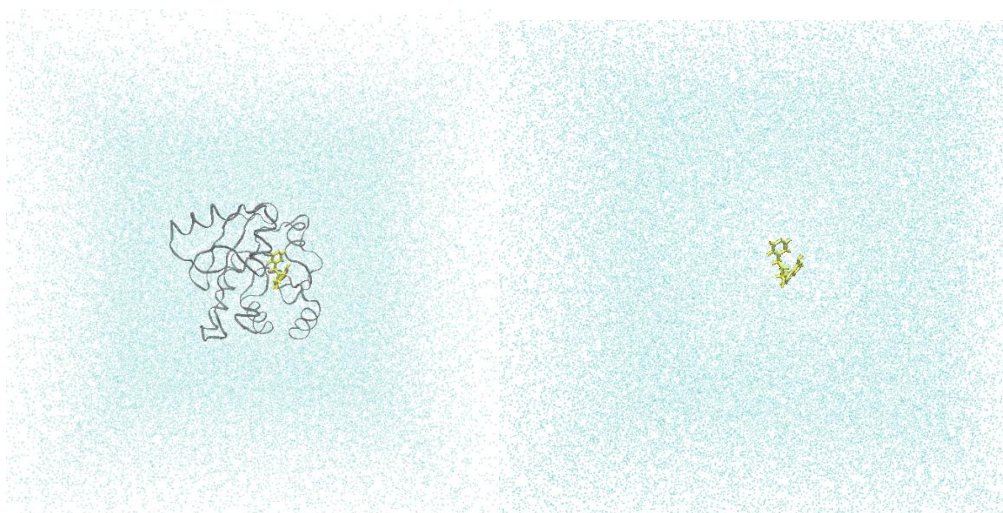
Lastly, for the first round of ABFE run on **CACHE_1689_35** (Figure 4.28. a), the free energies of the bound state and unbound states were 27.34 kcal/mol and 16.87 kcal/mol, respectively (Figure 4.28. b, c). A standard-state correction of approximately 6.9 kcal/mol was added to account for the Boresch restraints which was calculated by the MDRestraintsGenerator. After computing these values to the following equation, the ΔG_{pred}^0 binding free energy value was -3.57 kcal/mol.

$$\Delta G_{pred}^0 = -27.34 + 16.87 + 6.9$$

$$\Delta G_{pred}^0 = -3.57 \text{ kcal/mol}$$

The predicted ΔG_{pred}^0 underestimates the binding free energy compared to the experimental ΔG_{exp}^0 i.e., -7.3 kcal/mol. The ΔG_{pred}^0 binding free energy values of second and the third runs were -3.69 kcal/mol and -4.13 kcal/mol, respectively (**Appendix 6.8.**). The coulombic decoupling in the bound state showed a smooth, gradual trend in the free energy differences across the λ windows, while the van der Waals decoupling followed a bell-shaped curve (Figure 4.28. d). The results from the ABFE calculations of the three systems have been summarised in Table 4.1. As per the ΔG_{pred}^0 values, the ABFE workflow was able to identify S6U (5SPT) to be a better binder than RWQ (5SOJ).

(a)



(b)

Final results in kJ/mol:					Final results in kJ/mol:				
point	0 -	1,	DG	17.76 +/- 0.05	point	0 -	1,	DG	0.04 +/- 0.00
point	1 -	2,	DG	15.56 +/- 0.05	point	1 -	2,	DG	0.06 +/- 0.00
point	2 -	3,	DG	13.38 +/- 0.07	point	2 -	3,	DG	0.09 +/- 0.00
point	3 -	4,	DG	11.17 +/- 0.06	point	3 -	4,	DG	0.09 +/- 0.00
point	4 -	5,	DG	9.20 +/- 0.03	point	4 -	5,	DG	0.09 +/- 0.00
point	5 -	6,	DG	7.63 +/- 0.02	point	5 -	6,	DG	0.18 +/- 0.01
point	6 -	7,	DG	5.93 +/- 0.06	point	6 -	7,	DG	0.16 +/- 0.01
point	7 -	8,	DG	4.43 +/- 0.05	point	7 -	8,	DG	0.43 +/- 0.01
point	8 -	9,	DG	3.18 +/- 0.04	point	8 -	9,	DG	0.38 +/- 0.01
point	9 -	10,	DG	1.79 +/- 0.08	point	9 -	10,	DG	0.56 +/- 0.01
total	0 -	10,	DG	90.05 +/- 0.25	total	0 -	11,	DG	0.49 +/- 0.01
Bound_coul = 21.52 kcal/mol					Bound_restr = 0.61 kcal/mol				
Complex = 27.34 kcal/mol									

Final results in kJ/mol:

point	0 -	1,	DG	5.73 +/- 0.03
point	1 -	2,	DG	5.60 +/- 0.05
point	2 -	3,	DG	5.34 +/- 0.05
point	3 -	4,	DG	5.20 +/- 0.04
point	4 -	5,	DG	5.22 +/- 0.04
point	5 -	6,	DG	5.23 +/- 0.04
point	6 -	7,	DG	4.97 +/- 0.04
point	7 -	8,	DG	4.59 +/- 0.06
point	8 -	9,	DG	4.28 +/- 0.06
point	9 -	10,	DG	4.01 +/- 0.04
point	10 -	11,	DG	3.59 +/- 0.04
point	11 -	12,	DG	2.86 +/- 0.05
point	12 -	13,	DG	1.61 +/- 0.18
point	13 -	14,	DG	-0.85 +/- 0.44
point	14 -	15,	DG	-4.66 +/- 0.54
point	15 -	16,	DG	-8.43 +/- 0.35
point	16 -	17,	DG	-9.71 +/- 0.14
point	17 -	18,	DG	-7.70 +/- 0.06
point	18 -	19,	DG	-4.29 +/- 0.05
point	19 -	20,	DG	-0.79 +/- 0.06
total	0 -	20,	DG	21.81 +/- 1.24

Final results in kJ/mol:

point	0 -	1,	DG	0.04 +/- 0.00
point	1 -	2,	DG	0.06 +/- 0.00
point	2 -	3,	DG	0.09 +/- 0.00
point	3 -	4,	DG	0.09 +/- 0.00
point	4 -	5,	DG	0.09 +/- 0.00
point	5 -	6,	DG	0.18 +/- 0.01
point	6 -	7,	DG	0.16 +/- 0.01
point	7 -	8,	DG	0.43 +/- 0.01
point	8 -	9,	DG	0.38 +/- 0.01
point	9 -	10,	DG	0.56 +/- 0.01
point	10 -	11,	DG	0.49 +/- 0.01

Bound_restr = 0.61 kcal/mol

Bound_vdw = 5.21 kcal/mol

(c)

Final results in kJ/mol:					Final results in kJ/mol:				
point	0 -	1,	DG	16.08 +/- 0.06	point	0 -	1,	DG	4.51 +/- 0.03
point	1 -	2,	DG	13.40 +/- 0.11	point	1 -	2,	DG	4.44 +/- 0.03
point	2 -	3,	DG	10.98 +/- 0.07	point	2 -	3,	DG	4.34 +/- 0.04
point	3 -	4,	DG	9.02 +/- 0.05	point	3 -	4,	DG	4.22 +/- 0.02
point	4 -	5,	DG	7.12 +/- 0.07	point	4 -	5,	DG	4.04 +/- 0.03
point	5 -	6,	DG	5.50 +/- 0.05	point	5 -	6,	DG	3.89 +/- 0.01
point	6 -	7,	DG	4.02 +/- 0.03	point	6 -	7,	DG	3.70 +/- 0.01
point	7 -	8,	DG	2.74 +/- 0.05	point	7 -	8,	DG	3.43 +/- 0.02
point	8 -	9,	DG	1.71 +/- 0.04	point	8 -	9,	DG	3.11 +/- 0.02
point	9 -	10,	DG	0.59 +/- 0.03	point	9 -	10,	DG	2.69 +/- 0.02
total	0 -	10,	DG	71.15 +/- 0.14	point	10 -	11,	DG	2.09 +/- 0.01
Unbound_coul = 17.0 kcal/mol					point	11 -	12,	DG	1.34 +/- 0.02
Ligand in solvent = 16.87 kcal/mol					point	12 -	13,	DG	0.26 +/- 0.02
					point	13 -	14,	DG	-1.43 +/- 0.02
					point	14 -	15,	DG	-4.67 +/- 0.05
					point	15 -	16,	DG	-9.45 +/- 0.06
					point	16 -	17,	DG	-11.11 +/- 0.06
					point	17 -	18,	DG	-8.90 +/- 0.04
					point	18 -	19,	DG	-5.31 +/- 0.01
					point	19 -	20,	DG	-1.72 +/- 0.01
					total	0 -	20,	DG	-0.52 +/- 0.18
					Unbound_vdw = -0.124 kcal/mol				

(d)

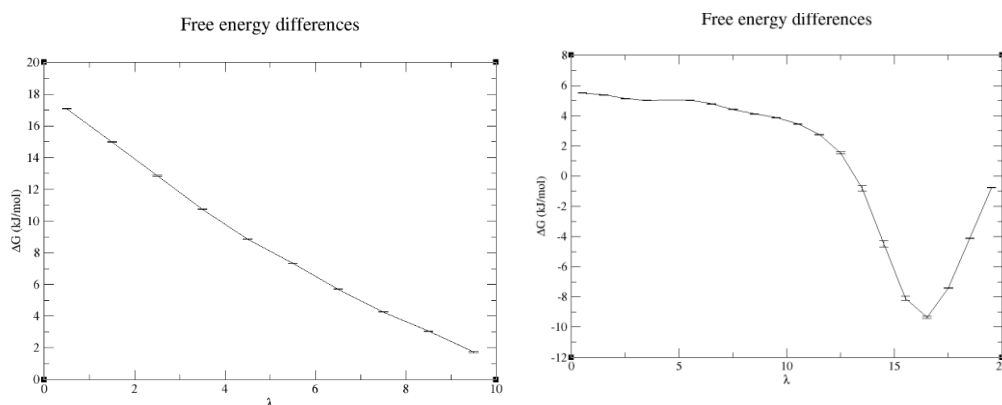


Figure 4.28. (a) Bound and unbound simulation box showing complex (CACHE_1689_35) in solvent and ligand in solvent (SK2) respectively. The ligand is shown in yellow; the protein is shown in grey, and the water molecules are shown in blue colour. Snapshot of the bound state free energy differences across the various windows of electrostatic (11 windows) and vdw decoupling (21 windows) along with restraints contribution (12 windows). (c) Snapshot of the unbound state free energy differences across the various windows of electrostatic (11 windows) and vdw decoupling (21 windows). The free energy difference (ΔG - ΔG) between two consecutive windows are calculated in kJ/mol. The cumulative free energy (kcal/mol) is calculated by adding the contribution of each point (2 consecutive windows) for each decoupling state. (d) Graphs showing trend in the free energy differences across the λ windows upon coulombic and vdw decoupling. The x-axis shows λ windows and y-axis shows free energy difference (kcal/mol).

System	Ligand	K_i/K_d (μM)	ΔG_{exp}^0	ΔG_{pred}^0 (1 run)	ΔG_{pred}^0 (2 run)	ΔG_{pred}^0 (3 run)	Average ΔG_{pred}^0	RMSE $Avg G_{pred}^0 - G_{exp}^0$
5SPT	S6U	52.3 μM	-5.84 kcal/mol	-6.84 kcal/mol	-6.79 kcal/mol	-6.80 kcal/mol	-6.81 kcal/mol	0.97 kcal/mol
5SOJ	RWQ	78.6 μM	-5.5 kcal/mol	-2.8 kcal/mol	-2.8 kcal/mol	-2.24 kcal/mol	-2.61 kcal/mol	2.89 kcal/mol
CACHE_1689_35	SK2	4.8 μM	-7.3 kcal/mol	-3.57 kcal/mol	-3.69 kcal/mol	-4.13 kcal/mol	-3.79 kcal/mol	3.51 kcal/mol

Table 4.1. Summary of ABFE calculations on the three systems. The experimental inhibition or dissociation constants (K_i μM or K_d μM) along with their corresponding experimental binding free energies (ΔG_{exp}^0) are mentioned in kcal/mol. The predicted binding free energies (ΔG_{pred}^0) were obtained from three independent simulation runs, and the average value (Average ΔG_{pred}^0) was calculated. The RMSE represents the deviation between the average predicted and experimental binding free energies in kcal/mol.

4.9. Discussion

MD simulations are widely applied in CADD and protein design since they enable the study of protein–ligand and protein–protein stability under dynamic, physiologically relevant conditions. The idea behind MD simulations is to simulate the real-life, dynamic behaviour of molecules using physics-based methods, but to carry out these simulations, a proper understanding of force fields is required, which are used to calculate the system's parameters for molecular modelling. The CHARMM36m and CGenFF force fields were used in the study to calculate the parameters of the protein and small molecule, respectively. These were used as they provide penalty scores to the parameters which indicate the correctness of the assigned parameter. The penalties lower than 10 indicate the parameter

is fair; the penalties between 10 and 50 mean some basic validation is recommended; penalties higher than 50 indicate a poor parameter and mandate extensive validation and optimisation. The two hits SK1 and SK2 had acceptable penalties, but the two positive controls had higher penalties that required QM/MM optimisation. To optimise the parameters, a tool called FFParm (Kumar & MacKerell, 2024) was selected. The quantum mechanics optimisation was performed using Psi4 to compute the CHARMM parameters. The workflow was very complex, and a lot of errors were encountered at the AutoFit:chargeAlpThol stage which led to us halting this avenue of investigation. In such cases, it is recommended to use other force fields to calculate the parameters or try other QM/MM optimisation tools. Due to time constraints, this could not be incorporated in our study. Despite the high penalty scores indicating reduced parameter accuracy, the ligands exhibited stable behaviour throughout the MD simulations.

For performing MD simulations there are many tools available, but GROMACS was selected for the study as it is open source and has well-established tutorials and an active public user forum. All the MD simulations and ABFE calculations were performed on Vimes and Rincewind MDG servers, since 100 ns MD runs were completed in the shortest wall-clock time in the benchmarking study. The MD simulations were run with GPU acceleration and 30 CPU threads. The most important step while preparing the MD input files was to select the solvent box of the correct size and solute-box edge padding. This is done to avoid periodic-image artefacts where, under the periodic boundary conditions (infinite copies), the complex/ligand in solvent can interact with its own images across the boundary if the box is too small. To avoid this, a cubic solvent box with 2 nm padding for the complex in solvent and 2.5 nm padding for the ligand in solvent was used. A larger box size for the ligand was selected as the output of the MD run is an input to ABFE calculations, and with decoupling of the interactions of the ligand in the solvent, there is a chance that if the box size is small, the ligand can see its own periodic image that would ultimately impact the free energy. The other important aspect is to consider the correct starting pose before moving forward with the MD simulation, as this can have an impact on the binding free energy calculations. For example, multiple diverse docked poses can be selected for MD simulations to further understand the interactions and binding modes (Liu & Kokubo et al., 2020). In this study, only a single top FRED docked pose of the two hits (SK1 and SK2) was utilised for the MD simulations, which represents a limitation of the research.

The RMSD plot of the protein-SK2 complex between the starting pose and the 100 ns MD simulation frames showed a stable trajectory. The average RMSD of the ligand SK2 was 2.5 Å with a few fluctuations, which indicates conformational changes. To check if the fluctuations were present even after 100 ns run, an additional 500 ns MD simulation was performed. The 500 ns MD simulation also showed that the ligand stays within the binding pocket but undergoes continuous conformational changes with an average RMSD of 3.8 Å. The hydrogen bond analysis on the protein-SK2 complex revealed that the hydrogen bond with the LYS53 was not maintained during the 100 ns MD simulation. This is the reason why the ligand SK2 shows spikes in the RMSD plots; the absence of the hydrogen bond allows the ligand to acquire different conformations. On the other hand, the hydrogen bonds formed with ASP20 were maintained for 87.86% of the MD run, and with ILE21, they were maintained for 93.09% of the MD run. The hydrogen bonds with ASP20 and ILE21 keep the ligand SK2 within the binding pocket despite the absence of a hydrogen bond with LYS53. The structural overlay between the MD representative frame and the docked pose showed an RMSD of 2.1 Å between the proteins and an RMSD of 0.97 Å between the ligand poses. This showed that both MD and docking predicted the same binding mode of SK2 with SC-2 Mac1-NSP3 protein. Lastly, the structural overlay between the 100 ns MD representative frame and the docked pose had an RMSD of 2.1 Å between the proteins and an RMSD of 0.97 Å between the ligand poses, which reflects a similar binding mode predicted by both MD simulation and docking.

The RMSD plot of the protein-SK1 complex between the starting pose and the MD simulation frames showed a high level of fluctuations, which indicates unbinding of the ligand. The protein chain and protein backbone showed a stable RMSD throughout the 100 ns run, but the ligand (SK1) left the binding pocket at approximately 20 ns. Since the MD trajectory showed the unbinding of SK1, it was not selected for further ABFE analysis. Three more docked poses of protein-SK1 with different orientations (Figure 4.29.) were subjected to MD simulations, but a similar trend was observed where SK1 undergoes unbinding after a few frames. In such a case, different docking algorithms can be utilised to predict a consensus pose, which can then be subjected to MD simulation.

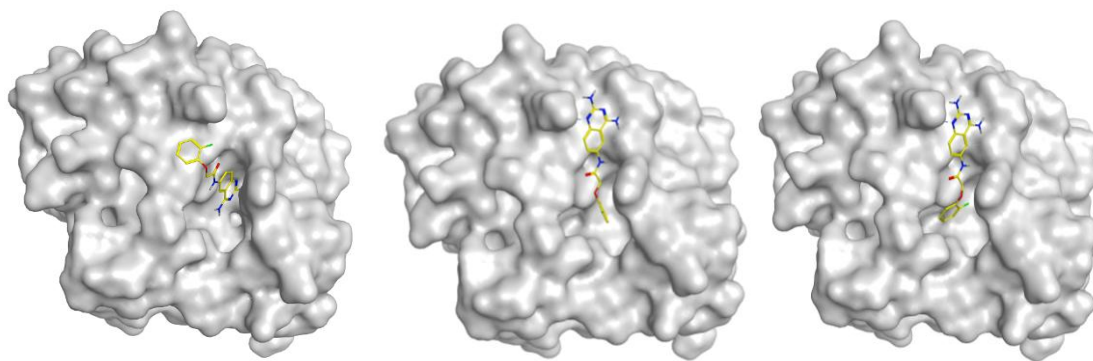
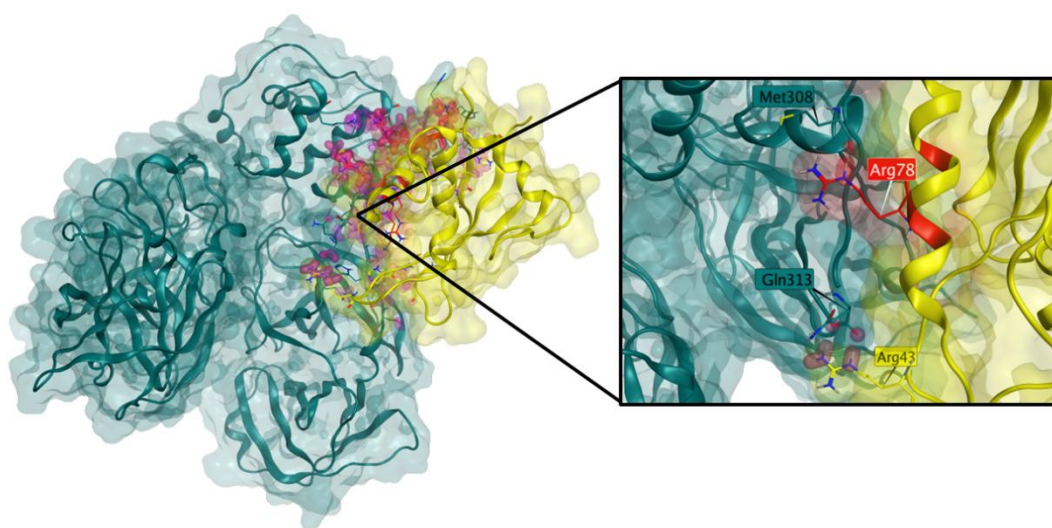


Figure 4.29. Three different docked poses of SK1 (yellow) with different orientations.

In a separate project with collaborators, we used the above-mentioned technique of using a consensus docking approach followed by MD simulations to model protein-protein interactions. A consensus docking approach with ClusPro and MOE was used to identify binding poses between the proteins, MERS-CoV NSP5 with IRF3 (Figure 4.30. a) and MERS-CoV-NSP5 with KPNA4 (Figure 4.30. b). In the study, there was biological evidence that MERS-CoV-NSP5 attenuates type I interferon production by inhibiting the nuclear translocation of IRF3. It was hypothesised that MERS-CoV-NSP5 might affect the translocation of IRF3 by interacting with KPNA4 importin by masking the nuclear localisation signal (NLS) region that interacts with IRF3. Molecular modelling was performed to observe the interactions between these proteins. The two docking algorithms predicted a similar binding pose for both MERS-CoV NSP5 with IRF3 and MERS-CoV-NSP5 with KPNA4. The binding poses were then subjected to 100 ns MD simulation using GROMACS. The structural overlays between the representative frames from the MD runs and the docked poses showed similar binding modes (Zhang et al., 2024). The results showed the masking of the NLS region on KPNA4 by MERS-CoV-NSP5, which thus inhibits the nuclear translocation of IRF3. It was also found that MERS-CoV-NSP5 interacted with IRF3 by masking the NLS region, potentially impacting its nuclear translocation.

(a)



(b)

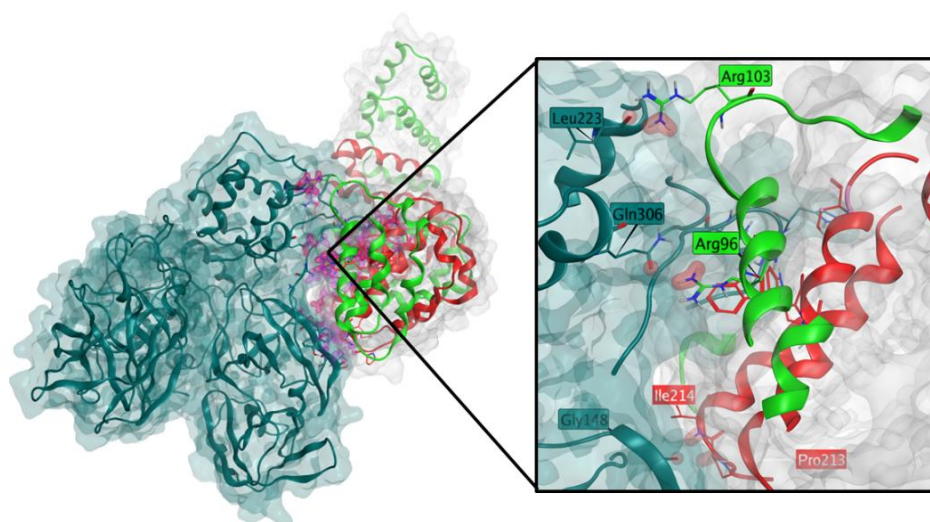


Figure 4.30. (a) Predicted binding mode of MERS-CoV NSP5 (dark green) with IRF3 (yellow). (b) Predicted binding mode of MERS-CoV NSP5 (dark green) with KPNA4 (light green). NLS region is highlighted in red. Schematic taken from (Zhang et al., 2024).

To predict the binding free energies between the protein and the ligand, there are three types of methods. The first method is the physics-based approach, which includes free energy perturbation (FEP) and thermodynamic integration (TI). This method utilises free-energy differences from a thermodynamic cycle by sampling the end states and multiple intermediate λ -windows to calculate the binding free energy. This approach is costlier and more time-consuming but typically more accurate than docking or end-state-based binding

free energies. The second approach is faster than FEP or TI, which includes the Molecular Mechanics Poisson–Boltzmann/Surface Area (MM-PBSA) and Molecular Mechanics–Generalised Born/Surface Area (MM-GBSA) (Rastelli et al., 2010; Kollman et al., 2000) methods, which rely on only sampling the end states (bound and unbound) and not the alchemical intermediate states. The last method is the machine learning (ML)-based methods, which consist of models that are trained on large datasets of experimentally measured binding affinities. For example, Boltz-2 (Passaro et al., 2025) is the first artificial intelligence (AI) model that approaches FEP-level performance for small-molecule affinity prediction and is faster than the physics-based method. As per the benchmarking study conducted by Passaro et al., 2025, traditional physics-based ABFE still outperforms in terms of accuracy, whereas Boltz-2 outperforms in terms of speed. Initially, to distinguish binders from non-binders, the MM-PBSA method was used by utilising a tool that works on GROMACS MD simulation output files called gmx_MMPBSA (Valdés-Tresanco et al., 2021). Three SC-2 Mac1-NSP3 x-ray crystal structures co-crystallised with lead-like molecules were selected, and their K_i values were determined. The three structures were 5SPB ($K_i = -8.78$ kcal/mol), 5SQA ($K_i = -7.71$ kcal/mol) and 5SRY ($K_i = -7.99$ kcal/mol). It was observed that gmx_MMPBSA overestimated the binding affinities (Table 4.2.). To rule out that this was due to the forcefield used, the ligands were also parameterised using the Amber forcefield, which also showed overestimated binding affinities. The MD output files from different run times (1 ns, 10 ns, 20 ns, 100 ns) were also tested, and they showed overestimated binding affinities. Since the MM-PBSA calculations produced overestimated binding affinities that did not correlate well with experimental binding affinity, we incorporated alchemical absolute binding free energy (ABFE) calculations using GROMACS to obtain more reliable estimates.

PDB	Experimental delta G	MMPBSA delta G Amber ff (10ns)	MMPBSA delta G CHARMM ff (10ns)	MMPBSA delta G CHARMM ff (1ns)	MMPBSA delta G CHARMM ff (20ns)	MMPBSA delta G CHARMM ff (100ns)
5SPB	-8.78 kcal/mol	-44.57 kcal/mol	-25 kcal/mol	-26.74 kcal/mol	-26.96 kcal/mol	-25.88 kcal/mol

5SQA	-7.71 kcal/mol	-31.76 kcal/mol	-16.63 kcal/mol	-17.86 kcal/mol	-17.43 kcal/mol	-16.60 kcal/mol
5SRV	-7.99 kcal/mol	-33.46 kcal/mol	-18.66 kcal/mol	-18.50 kcal/mol	-18.78 kcal/mol	-18.54 kcal/mol

Table 4.2. Comparison of predicted binding affinities (kcal/mol) between gmx_MMPBSA for CHARMM (green) and AMBER (orange) parametrised complexes with experimental binding affinities. For CHARMM-parameterised complexes, different MD run times were also used to observe any differences in binding affinities.

The ABFE calculations rely on proper selection of Boresch restraints, which help keep the ligand within the binding pocket during decoupling events. In this study MDRestrainsGenerator was used to automatically pick the ligand and protein atoms for generating the harmonic restraints, which was quicker than selecting the atoms manually. For the manual selection of protein and ligand atoms, MOE (2022.02) was used to calculate the distance, dihedral, and angle between them in the PDB file extracted from the last frame of the MD simulation output. Different sets of the atoms were selected to check if they were stable while running ABFE calculations which was very time consuming. The atom selection with MDRestrainsGenerator helped in selecting the most stable atoms to act as anchor points for the restraints. The use of MDRestrainsGenerator significantly improved the predicted binding affinities, whereas with the manual selection of atoms, the values were underestimated.

The ABFE replicate run on SC-2 Mpro in complex with Z219104216 helped in validating the capacity of GROMACS to carry out an ABFE calculation. The ABFE calculations for the two SC-2 Mac1-NSP3 positive control systems and CACHE_1689_35 included a total of 11 λ windows for electrostatic decoupling, 21 λ windows for vdw decoupling, and 12 λ windows to estimate the contribution of restraints. Each window was an independent MD simulation with energy minimisation, NVT equilibration (with position restraints), NPT equilibration (with position restraints), a second NPT equilibration (without position restraints), and finally MD production. For each window simulation, configurations were sampled at every 0.2 ps. For the ABFE workflow an extra round of NPT without the restraints was added

which was a different workflow than that of the validation ABFE workflow by Ghahremanpour et al., 2023. The additional second NPT equilibration (without position restraints) was added into the workflow to allow natural relaxation before the MD production (Alibay et al., 2022).

As per the ΔG_{pred}^0 values (S6U $\Delta G_{pred}^0 = -6.81$ kcal/mol, RWQ $\Delta G_{pred}^0 = -2.61$ kcal/mol), the ABFE workflow was able to identify S6U (5SPT, $\Delta G_{exp}^0 = -5.84$ kcal/mol) to be a better binder than RWQ (5SOJ, $\Delta G_{exp}^0 = -5.51$ kcal/mol). With regards to the CACHE_1689_35, the predicted average ΔG_{pred}^0 of -3.79 kcal/mol underestimates the binding free energy compared to the experimental ΔG_{exp}^0 i.e., -7.3 kcal/mol. To rule out sampling issues, the number of windows and the MD production run time were increased. It was found that these changes did not improve the ΔG_{pred}^0 . Several factors may have contributed to the overestimation or underestimation of the binding free energies. Firstly, the $K_d = 4.8$ μ M of CACHE_1689_35 was found in the first SPR run with a low binding percentage, hence, the experimental ΔG would not represent the actual binding free energy value. The root mean square error (RMSE) between the experimental and predicted binding free energies varies from ~ 0.80 to ~ 3.80 kcal/mol (Ghahremanpour et al., 2023). Different studies have reported higher RMSEs, e.g., ~ 2.7 - 3.8 kcal/mol (Alibay et al., 2022) and 2.3 kcal/mol (Wang et al., 2006). In our study the RMSE varied from 0.97 to 3.51 kcal/mol. The second reason may be that the parameters generated by CGenFF for the small molecules require further QM/MM optimisation. This is the forcefield limitation where the parameters generated for the system by CHARMM were not sufficient to model the interactions. Hence, other forcefields can be utilised to model the interactions and compare the binding free energies.

4.10. Conclusions and future directions

- A working protocol was established to model protein-protein or protein-ligand interactions using MD simulations with the GROMACS tool.
- To calculate the absolute binding free energies of protein-ligand complexes, a working protocol was developed using GROMACS.

- The RMSEs between the experimental and predicted binding free energies varied between 0.97 and 3.51 kcal/mol, which are within the acceptable range as per the literature.
- The ABFE protocol can be used to identify binders from non-binders.
- Since the ABFE protocol involves many manual inputs, an automated workflow using GROMACS with different force field options is required.
- The SC-2 Mac1-NSP3 can be modelled using other force fields, and the ABFE values obtained can be compared to the values obtained by modelling with the CHARMM force field.
- To optimise the parameters for the small molecules, QM/MM optimisation tools can be integrated into the automated ABFE workflow.
- The MD and ABFE protocol can be replicated on the structural analogues of the two SC-2 Mac1-NSP3 hits.

Chapter V.

Conclusion

&

Perspectives

During the end of 2019, there were cases of patients suffering from fever, malaise, dry cough, and dyspnoea diagnosed as viral pneumonia in Wuhan, China (Huang et al., 2019). Later, it was reported that the causative agent is a novel coronavirus, SARS-CoV-2 (SC-2). The infection caused by the virus spread quickly globally and was known as the COVID-19 pandemic. The scientists across the globe targeted different SC-2 viral proteins for the development of a therapeutic alternative. But as time had progressed, the virus underwent various mutations, thereby resulting in drug resistance. One such example is of SC-2 viral RNA-dependent RNA polymerase (RdRp), which underwent mutations in the active site resulting in 38-fold resistance to Remdesivir (Stevens et al., 2022). This pointed out the importance of targeting and designing drug molecules for binding sites on the SC-2 proteins, like the non-structural proteins (NSPs) that do not mutate over time, thereby helping in minimizing the risk of drug resistance caused due to mutations. With the help of computer-aided drug design technologies, the process of identifying molecules that could target these sites can be accelerated. The work done in this thesis has successfully identified conserved binding sites on the SC-2 NSPs, and it has also identified interesting molecules that target conserved sites on NSP13 and NSP3 proteins by utilising bioinformatics, CADD, and binding assays.

The **chapter I** in the thesis focuses on exploring the literature around the classification of viruses, specifically focusing on the Group IV viruses, including the families *Coronaviridae* and *Togaviridae*. The chapter specifically focuses on the SC-2 virus for *Coronaviridae* and the Chikungunya virus (CHIKV) for *Togaviridae*. The chapter discusses the genomic organisation of both viruses, their mechanism of viral entry into host cells, and current treatment options. The chapter also discusses the key concepts of bioinformatics and CADD techniques used in the project, outlining how these computational disciplines have become essential tools for accelerating drug discovery. It gives an overall understanding of these viruses as well as the technologies employed in the study, and it provides foundational knowledge for the development of the computational pipeline used in the project. It also discusses the different NSPs present in the viruses, which helps in correlating their functional activities for the development of a common drug candidate that could be utilised to target them. The literature support in the chapter has shaped the following work by utilising information gathered on the viral families and the computational technologies. Lastly, the chapter also summarises the hypotheses and objectives of the project.

The **chapter II** in the thesis describes the bioinformatics work performed in the project. It first describes the Kalign software used for multiple sequence alignments (MSAs), then explains what conserved regions are, presents previously reported literature evidence of conservation across different RNA viruses, discusses conservation across coronaviruses, and finally covers NSP conservation across coronaviruses. The chapter explores the bioinformatics methodology to identify conserved amino acid residues in the binding pockets of SC-2 NSPs. This study was crucial because most previous research has focused only on the SC-2 spike, protease, and polymerase viral targets, however, this study provides an extensive analysis of SC-2 NSPs, which are essential for both viral replication and pathogenicity.

The bioinformatics analysis of SC-2 NSP sequences involved identifying conserved binding sites and top mutations/deletions by aligning ~1 million protein sequences (from the NCBI dataset). The initial benchmarking study also provided an insight into the capability of different MSA tools in aligning large sequence datasets. The large size of the dataset used in the analysis facilitated the identification of more populated mutated and conserved amino acid residues. The workflow also provides a step-by-step methodology to prepare and sanitise the protein sequence dataset prior to performing MSAs. The sequence analysis performed on each NSP provides an insight about the level of conservation in each protein and also highlights the presence of conserved binding pockets that could be targeted in future studies for the development of pan SC-2 inhibitors (since the dataset consisted of SC-2 variants). Similarly, the mutational analysis performed highlights the presence of highly variant amino acid positions as well as deletions, which can be studied further by conducting mutagenesis studies to understand how they improve the overall virus's fitness, aid immune evasion and affect protein stability. Furthermore, the key mutations identified in each NSP can be monitored to support the surveillance of emerging SC-2 variants.

This chapter provides bioinformatics support for the project's hypotheses by identifying conserved binding pockets. These pockets can subsequently be targeted for the development of suitable inhibitors through the application of CADD techniques. Overall, the bioinformatics pipeline mentioned in this thesis, starting with sequence alignments, sequence data processing, conservation/mutational analysis and 3D visualisations of conserved and mutated amino acid positions, can be replicated for different projects involving protein sequences.

The **chapter III** in the thesis first utilises the bioinformatics workflow, which is applied to two NSPs, like SC-2 NSP13 (RNA binding site) and the SC-2 Mac1 domain of the NSP3 (ADP-r binding site). The conservation of the binding sites of the two NSPs was not only limited to the large-scale SC-2 variant sequence dataset but also across the betacoronaviruses; these findings would thus be beneficial for the identification of a pan-coronavirus inhibitor. Additionally, the concept of sequence conservation was examined from a broader perspective, specifically investigating whether these two binding pockets were conserved in group IV viruses (+ssRNA), with a particular emphasis on the chikungunya virus. A high level of sequence as well as structural conservation was established between these viruses, which may represent novel starting points for designing small molecule interventions across both viruses. As for the binding sites of the two proteins, they are validated for the development of therapeutic interventions.

In order to identify novel small molecule inhibitors for these binding sites, the computational drug design approaches were utilised. The chapter also covers the CACHE challenges, which were focused on these two proteins. The computational workflow involved the virtual screening of the Enamine database using volume-based screening, pharmacophore-based screening, and molecular docking to identify potential small-molecule inhibitors. The hit lists obtained after the VS workflows were sent to CACHE, where the compounds were experimentally validated with binding assays. For the challenge #2 (SC-2 NSP13), after the completion of binding assays, five hits were identified but were not progressed to hit-to-lead optimisation. The NSP13 hits identified in this study may influence viral replication and pathogenicity, given the multifunctional role of this protein, but experimental studies will be necessary to determine the actual impact of these molecules. It will also be interesting to identify structural analogues of NSP13 hits and retest them in future studies. As for challenge #3 (SC-2 NSP3), two hits were identified after the completion of binding assays by CACHE, and a hit-to-lead optimisation study was carried out with the aid of the Notre Dame-DCU Seed Grant in Precision Biomedical Technologies funding. The analogues will be tested in binding assays against the SC-2 Mac1-NSP3 as well as the CHIKV Mac1 domain by Dr Choon Kim at Notre Dame University, USA. For future studies, the impact of these hits can be validated experimentally; for example, as the Mac1 domain plays a key role in counteracting host immune responses, its inhibition by these hits could potentially attenuate viral pathogenicity.

Overall, the hits identified in the two challenges provide promising starting points for further investigation. The conservation of the binding sites, i.e., SC-2 Mac1-NSP3 and SC-2 NSP13 (RNA binding site), across the betacoronaviruses (SC-2, SC-1, MERS-CoV) and CHIKV points to the possibility of discovering pan-antiviral small molecule inhibitors which will offer protection from possible future outbreaks caused by related viruses. The CADD protocol developed in this project is available for early-stage drug discovery initiatives focused on any protein of interest. Additionally, integrating the bioinformatics pipeline discussed in chapter II can assist in pinpointing the most conserved binding sites and recognising significant mutations.

The **chapter IV** covers the methodology and outcomes of much more rigorous computational techniques involving molecular dynamics and absolute binding free energy calculations. It covers the key concepts behind protein-ligand binding, considering all-atom dynamics and the thermodynamic cycle to understand ligand decoupling. Firstly, a working protocol to model protein-ligand interactions using MD simulations with GROMACS software is implemented. Secondly, to the best of our knowledge, the working ABFE protocol developed in this project is the first on the SC-2 NSP3 Mac1 domain protein to predict the binding free energies of protein-ligand systems. The protocols for both MD and ABFE simulations were validated by performing benchmarking runs on known binders (active molecules). Since ABFE simulations are the best way to identify and rank binders from non-binders, the protocol can be implemented on the structural analogues of the hits identified in chapter III. As mentioned, the ABFE protocol is a rigorous simulation requiring expertise and manual inputs; an automated workflow using GROMACS with different force field options can be implemented in future studies. Additionally, QM/MM optimisation tools can also be integrated into the automated ABFE workflow to optimise the parameters for small molecules. Currently, even though there are several protocols that use ML and AI to predict the binding free energies of protein-ligand complexes, the traditional physics-based or simulation-driven approaches still prove more reliable in many instances. For example, the ML-based tool Boltz-2 demonstrates potential in terms of speed and scalability, but as it relies frequently on training data, it struggles with the estimation when applied to unfamiliar systems (Passaro et al., 2025).

In conclusion, the work presented in the thesis provides a state-of-the-art working protocol involving bioinformatics analysis, computer-aided drug design and rigorous computational

simulations. The ultimate aim of the project was to identify a common key for a common lock, the key being the hit molecules and the lock representing the conserved binding pockets across NSPs of SC-2 and related viruses. A combination of the computational strategies has resulted in hits (starting points) that could be further explored to tackle future outbreaks that might be caused by these viruses. A unified CADD platform can be developed in subsequent studies utilising the computational methods mentioned in this project to support academic groups and pharmaceutical companies with early-stage drug discovery projects.

286 | Page

Appendix 6.2. Bash commands for SC-2 variant analysis

Alpha

```
>>> grep -A 1 -E '^>(. *B\.\.1\.\.1\.\.7$|Q\.[0-9]+))' one_line_alpha.fasta > alpha_test.fasta and
>>> grep -A 1 -E '^>.*B\.\.1\.\.1\.\.7[0-9]*$' one_line_alpha.fasta |
> grep -v '^--$' > only_alpha.fasta
```

Beta

```
>>> grep -A 1 -E '^>.*B\.\.1\.\.351(\.[0-9]+)?$' one_line_beta.fasta |
> grep -v '^--$' > only_beta_sequences.fasta
```

Delta

```
>>> grep -A 1 -E '^>.*AY\.'" one_line_delta.fasta | grep -v '^--$' > only_delta1.fasta
>>> grep -A 1 -E '^>.*\|B\.\.1\.\.617\.\.2$' one_line_delta.fasta |
> grep -v '^--$' > only_delta2.fasta
>>> grep -A 1 -E '^>.*\|XS$' one_line_delta.fasta | grep -v '^--$' > only_delta3.fasta
>>> awk -v RS=">" '!/AY\./{print ">"$0}' without_alpha_beta.fasta >
without_alpha_beta_AYdelta.fasta (without AY)
>>> sed -i '/^$/d' without_alpha_beta_AYdelta.fasta
>>> grep -v -E '^>.*\|B\.\.1\.\.617\.\.2$' without_alpha_beta_AYdelta.fasta | grep -v '^--$' >
without_alpha_beta_B.1.617.2_delta.fasta (without B.1.617.2)
```

Gamma

```
>>> grep -A 1 -E '^>.*P\.'" one_line_gamma.fasta | grep -v '^--$' > only_gammaP1.fasta
```

Omicron

```
>>> chmod +x extract_sequences_2.bash
>>> ./extract_sequences_2.bash
>>> cat *fasta > all_omicron.fasta
>>> chmod +x non_header_pattern.sh
>>> ./non_header_pattern.sh (without omicron)
```

Non-VOC**Alpha**

```
>>> awk -v RS=">" '!/B\1\1\7(\.[0-9]+)?/{print ">"$0}' one_line_alpha.fasta >
1_without_alpha.fasta
```

```
>>> sed -i '/^$/d' 1_without_alpha.fasta
```

Beta

```
>>> awk -v RS=">" '!/B\1\351(\.[0-9]+)?/{print ">"$0}' 1_without_alpha.fasta >
2_without_alpha_beta.fasta
```

```
>>> sed -i '/^$/d' 2_without_alpha_beta.fasta
```

Delta

```
>>> awk -v RS=">" '!/AY\./{print ">"$0}' 2_without_alpha_beta.fasta >
3_without_alpha_beta_AY_delta.fasta
```

```
>>> sed -i '/^$/d' 3_without_alpha_beta_AY_delta.fasta
```

```
>>> awk -v RS=">" '!/B\1\617\2/{print ">"$0}' 3_without_alpha_beta_AY_delta.fasta >
3_2_without_alpha_beta_AY_delta_B16172.fasta
```

```
>>> sed -i '/^$/d' 3_2_without_alpha_beta_AY_delta_B16172.fasta
```

```
>>> grep -A 1 -E '^>.*[^\X]$$/^>.*X[^\S]$/^>.*[^\XS]$$'
```

```
4_2_without_alpha_beta_gamm_AY_B.617_delta.fasta >
```

```
4_3_without_alpha_beta_gamm_AY_B.617_XS_delta.fasta
```

```
>>> cat *fasta > all_delta.fasta
```

Gamma

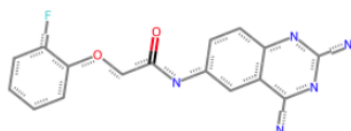
```
>>> awk -v RS=">" '!/P\./{print ">"$0}' 3_3_without_alpha_beta_AY_delta_B16172_XS.fasta >
4_without_alpha_beta_delta_gamma.fasta
```

```
>>> sed -i '/^$/d' 4_without_alpha_beta_delta_gamma.fasta
```

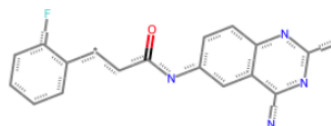
Appendix 6.3. Substructure search of CACHE hits

6.3.1. Substructure search batches of CACHE3HI_1689_15

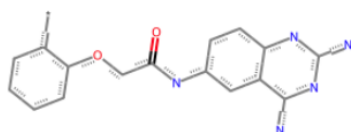
Pattern Name: original
 SMARTS: Fc1c(OCC(=O)Nc2cc3c(N)nc(N)nc3cc2)cccc1



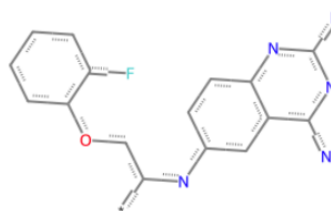
Pattern Name: with_any_hydroxyl
 SMARTS: Fc1c(*CC(=O)Nc2cc3c(N)nc(N)nc3cc2)cccc1



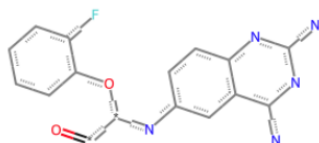
Pattern Name: with_any_F
 SMARTS: *c1c(OCC(=O)Nc2cc3c(N)nc(N)nc3cc2)cccc1



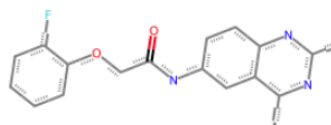
Pattern Name: with_any_acetyl
 SMARTS: Fc1c(OCC(*)Nc2cc3c(N)nc(N)nc3cc2)cccc1



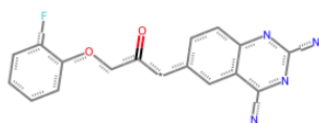
Pattern Name: with_any_linker
 SMARTS: Fc1c(O*(*)Nc2cc3c(N)nc(N)nc3cc2)cccc1



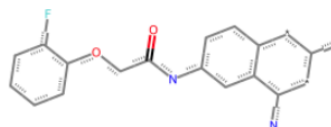
Pattern Name: with_any_two_nh2
 SMARTS: Fc1c(OCC(=O)Nc2cc3c(*)nc(*)nc3cc2)cccc1



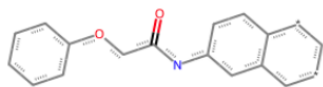
Pattern Name: with_any_amide
 SMARTS: Fc1c(OCC(=O)*c2cc3c(N)nc(N)nc3cc2)cccc1



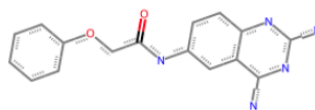
Pattern Name: with_any_ring_N
 SMARTS: Fc1c(OCC(=O)Nc2cc3c(N)*c(N)*c3cc2)cccc1



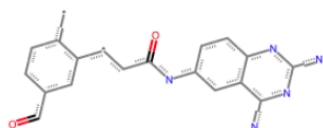
Pattern Name: without_F_and_nh2_with_any_ring_N
 SMARTS: c1c(OCC(=O)Nc2cc3c*c*c3cc2)cccc1



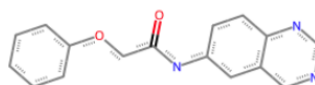
Pattern Name: without_F
 SMARTS: c1c(OCC(=O)Nc2cc3c(N)nc(N)nc3cc2)cccc1



Pattern Name: with_extended_HBA_any_F_any_hydroxyl
 SMARTS: *c1c(*CC(=O)Nc2cc3c(N)nc(N)nc3cc2)cc(C=O)cc1

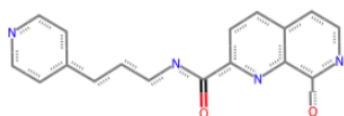


Pattern Name: without_F_and_nh2
 SMARTS: c1c(OCC(=O)Nc2cc3cncnc3cc2)cccc1

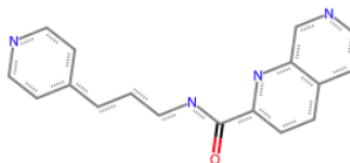


6.3.2. Substructure search batches of CACHE3HI_1689_35

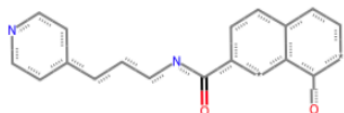
Pattern Name: with_o
 SMARTS: O=C(NCCCC1ccncc1)c1nc2c(O)nccc2cc1



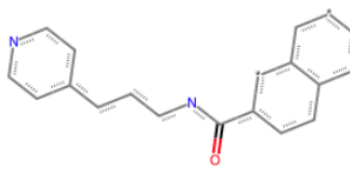
Pattern Name: without_o
 SMARTS: O=C(NCCCC1ccncc1)c1nc2c(cc1)ccnc2



Pattern Name: with_o_any_n
 SMARTS: O=C(NCCCC1ccncc1)c1*c2c(O)*ccc2cc1

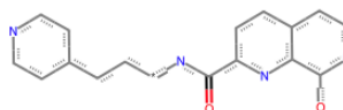
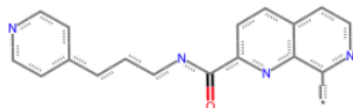


Pattern Name: without_o_any_n
 SMARTS: O=C(NCCCC1ccncc1)c1*c2c(cc1)cc*c2



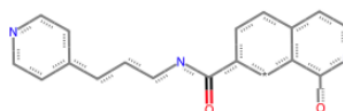
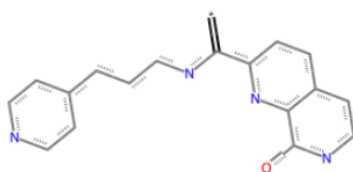
Pattern Name: with_any_o_but_all_N
 SMARTS: O=C(NCCCC1CNC1)C1NC2C(*)NCCC2CC1

Pattern Name: with_any_side_chain
 SMARTS: O=C(N*CC1CNC1)C1NC2C(O)NCCC2CC1

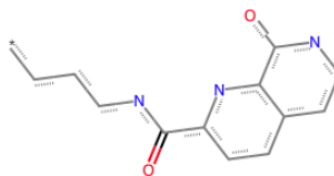


Pattern Name: With_any_oxo_but_all_N
 SMARTS: *=C(NCCCC1CNC1)C1NC2C(O)NCCC2CC1

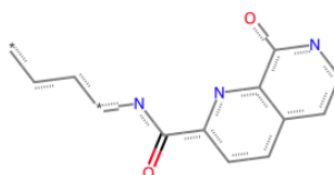
Pattern Name: with_any_on_pyridine_ring
 SMARTS: O=C(NCCCC1CNC1)C1*c2c(O)NCCC2CC1



Pattern Name: with_any_on_terminal_pyridine_rir
 SMARTS: O=C(NCCC*)C1NC2C(O)NCCC2CC1



Pattern Name: with_flexible_linker
 SMARTS: O=C(N*CC*)C1NC2C(O)NCCC2CC1



Appendix 6.4. GROMACS MD simulation MDP files

(i) ions.mdp

```
; LINES STARTING WITH ';' ARE COMMENTS
title           = Minimization ; Title of run

; Parameters describing what to do, when to stop and what to save
integrator      = steep        ; Algorithm (steep = steepest descent minimization)
emtol          = 1000.0        ; Stop minimization when the maximum force < 10.0 kJ/mol
emstep         = 0.01          ; Energy step size
nsteps         = 50000         ; Maximum number of (minimization) steps to perform

; Parameters describing how to find the neighbors of each atom and how to calculate the interactions
nstlist        = 1             ; Frequency to update the neighbor list and long range forces
cutoff-scheme   = Verlet
ns_type        = grid          ; Method to determine neighbor list (simple, grid)
rlist          = 1.0           ; Cut-off for making neighbor list (short range forces)
coulombtype     = cutoff       ; Treatment of long range electrostatic interactions
rcoulomb       = 1.0           ; long range electrostatic cut-off
rvdw           = 1.0           ; long range Van der Waals cut-off
pbc            = xyz           ; Periodic Boundary Conditions
```

(ii) em.mdp

```
; LINES STARTING WITH ';' ARE COMMENTS
title           = Minimization ; Title of run

; Parameters describing what to do, when to stop and what to save
integrator      = steep        ; Algorithm (steep = steepest descent minimization)
emtol          = 100.0         ; Stop minimization when the maximum force < 10.0 kJ/mol
emstep         = 0.01          ; Energy step size
nsteps         = 5000000       ; Maximum number of (minimization) steps to perform

; Parameters describing how to find the neighbors of each atom and how to calculate the interactions
nstlist        = 1             ; Frequency to update the neighbor list and long range forces
cutoff-scheme   = Verlet
ns_type        = grid          ; Method to determine neighbor list (simple, grid)
rlist          = 1.2           ; Cut-off for making neighbor list (short range forces)
coulombtype     = PME          ; Treatment of long range electrostatic interactions
rcoulomb       = 1.2           ; long range electrostatic cut-off
vdwtype        = cutoff
vdw-modifier    = force-switch
rvdw-switch     = 1.0
rvdw           = 1.2           ; long range Van der Waals cut-off
pbc            = xyz           ; Periodic Boundary Conditions
DispCorr       = no
```

(iii) nvt.mdp

```

title                = Protein-ligand complex NVT equilibration
define               = -DPOSRES ; position restrain the protein and ligand
; Run parameters
integrator           = md        ; leap-frog integrator
nsteps               = 500000    ; 2 * 50000 = 100 ps
dt                   = 0.002     ; 2 fs
; Output control
nstenergy            = 500       ; save energies every 1.0 ps
nstlog               = 500       ; update log file every 1.0 ps
nstxout-compressed   = 500       ; save coordinates every 1.0 ps
; Bond parameters
continuation         = no        ; first dynamics run
constraint_algorithm = lincs      ; holonomic constraints
constraints          = h-bonds   ; bonds to H are constrained
lincs_iter           = 1         ; accuracy of LINCS
lincs_order          = 4         ; also related to accuracy
; Neighbor searching and vdW
cutoff-scheme        = Verlet
ns_type              = grid      ; search neighboring grid cells
nstlist              = 20        ; largely irrelevant with Verlet
rlist                = 1.2
vdwtype              = cutoff
vdw-modifier          = force-switch
rvdw-switch          = 1.0
rvdw                 = 1.2       ; short-range van der Waals cutoff (in nm)
; Electrostatics
coulombtype          = PME        ; Particle Mesh Ewald for long-range electrostatics
rcoulomb             = 1.2       ; short-range electrostatic cutoff (in nm)
pme_order            = 4         ; cubic interpolation
fourierspacing       = 0.16     ; grid spacing for FFT
; Temperature coupling
tcoupl               = V-rescale   ; modified Berendsen thermostat
tc-grps              = Protein_Ligand Water_Ions ; two coupling groups - more accurate
tau_t                = 0.1 0.1    ; time constant, in ps
ref_t                = 300 300    ; reference temperature, one for each group, in K
; Pressure coupling
pcoupl               = no         ; no pressure coupling in NVT
; Periodic boundary conditions
pbc                  = xyz        ; 3-D PBC
; Dispersion correction is not used for proteins with the C36 additive FF
DispCorr             = no
; Velocity generation
gen_vel              = yes        ; assign velocities from Maxwell distribution
gen_temp             = 300        ; temperature for Maxwell distribution
gen_seed             = -1         ; generate a random seed

```


(iv) npt.mdp

```

title                = Protein-ligand complex NPT equilibration
define               = -DPOSRES ; position restrain the protein and ligand
; Run parameters
integrator           = md ; leap-frog integrator
nsteps               = 2500000 ; 2 * 50000 = 100 ps
dt                   = 0.002 ; 2 fs
; Output control
nstenergy            = 500 ; save energies every 1.0 ps
nstlog               = 500 ; update log file every 1.0 ps
nstxout-compressed   = 500 ; save coordinates every 1.0 ps
; Bond parameters
continuation         = yes ; continuing from NVT
constraint_algorithm = lincs ; holonomic constraints
constraints          = h-bonds ; bonds to H are constrained
lincs_iter           = 1 ; accuracy of LINCS
lincs_order          = 4 ; also related to accuracy
; Neighbor searching and vdW
cutoff-scheme        = Verlet
ns_type              = grid ; search neighboring grid cells
nstlist              = 20 ; largely irrelevant with Verlet
rlist                = 1.2
vdwtype              = cutoff
vdw-modifier          = force-switch
rvdw-switch          = 1.0
rvdw                 = 1.2 ; short-range van der Waals cutoff (in nm)
; Electrostatics
coulombtype          = PME ; Particle Mesh Ewald for long-range electrostatics
rcoulomb             = 1.2
pme_order            = 4 ; cubic interpolation
fourierspacing       = 0.16 ; grid spacing for FFT
; Temperature coupling
tcoupl               = V-rescale ; modified Berendsen thermostat
tc-grps              = Protein_Ligand Water_Ions ; two coupling groups - more accurate
tau_t                = 0.1 0.1 ; time constant, in ps
ref_t                = 300 300 ; reference temperature, one for each group, in K
; Pressure coupling
pcoupl               = Berendsen ; pressure coupling is on for NPT
pcoupltype           = isotropic ; uniform scaling of box vectors
tau_p                = 2.0 ; time constant, in ps
ref_p                = 1.0 ; reference pressure, in bar
compressibility       = 4.5e-5 ; isothermal compressibility of water, bar^-1
refcoord_scaling      = com
; Periodic boundary conditions
pbc                  = xyz ; 3-D PBC
; Dispersion correction is not used for proteins with the C36 additive FF
DispCorr             = no
; Velocity generation
gen vel              = no ; velocity generation off after NVT

```

(v) md.mdp

```

title                      = Protein-ligand complex MD simulation
; Run parameters
integrator                 = md          ; leap-frog integrator
nsteps                    = 50000000 ; 2 * 5000000 = 10000 ps (10 ns)
dt                        = 0.002      ; 2 fs
; Output control
nstenergy                 = 5000        ; save energies every 10.0 ps
nstlog                    = 5000        ; update log file every 10.0 ps
nstxout-compressed        = 5000        ; save coordinates every 10.0 ps
; Bond parameters
continuation              = yes         ; continuing from NPT
constraint_algorithm      = lincs       ; holonomic constraints
constraints               = h-bonds     ; bonds to H are constrained
lincs_iter                = 1           ; accuracy of LINCS
lincs_order               = 4           ; also related to accuracy
; Neighbor searching and vdW
cutoff-scheme             = Verlet
ns_type                   = grid        ; search neighboring grid cells
nstlist                   = 20          ; largely irrelevant with Verlet
rlist                     = 1.2
vdwtype                   = cutoff
vdw-modifier              = force-switch
rvdw-switch               = 1.0
rvdw                      = 1.2        ; short-range van der Waals cutoff (in nm)
; Electrostatics
coulombtype               = PME         ; Particle Mesh Ewald for long-range electrostatics
rcoulomb                  = 1.2
pme_order                 = 4           ; cubic interpolation
fourierspacing            = 0.16       ; grid spacing for FFT
; Temperature coupling
tcoupl                    = V-rescale   ; modified Berendsen thermostat
tc-grps                   = Protein_Ligand Water_Ions ; two coupling groups - more accurate
tau_t                     = 0.1 0.1    ; time constant, in ps
ref_t                     = 300 300    ; reference temperature, one for each group, in K
; Pressure coupling
pcoupl                    = Parrinello-Rahman ; pressure coupling is on for NPT
pcoupltype                = isotropic   ; uniform scaling of box vectors
tau_p                     = 2.0         ; time constant, in ps
ref_p                     = 1.0         ; reference pressure, in bar
compressibility            = 4.5e-5      ; isothermal compressibility of water, bar^-1
; Periodic boundary conditions
pbc                       = xyz         ; 3-D PBC
; Dispersion correction is not used for proteins with the C36 additive FF
DispCorr                  = no
; Velocity generation
gen_vel                   = no          ; continuing from NPT equilibration

```

Appendix 6.5. GROMACS ABFE window spacing

1. Coulombic decoupling 11 windows (1: interactions off, 0: interactions on)

bonded-lambdas = 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0

coul-lambdas = 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

vdw-lambdas = 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

2. Vdw decoupling 21 windows

bonded-lambdas= 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0

coul-lambda = 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0 1.00 1.0

vdw-lambdas = 0.0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1.0

3. Restraints contribution 12 windows

bonded-lambdas = 0.0 0.01 0.025 0.05 0.075 0.1 0.15 0.2 0.35 0.5 0.75 1.0

Appendix 6.6. Second and third ABFE runs on 5SPT

(i) Run 2

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 23.20 +/- 0.11	point	0 -	1,	DG 0.03 +/- 0.00
point	1 -	2,	DG 21.11 +/- 0.09	point	1 -	2,	DG 0.04 +/- 0.00
point	2 -	3,	DG 19.28 +/- 0.07	point	2 -	3,	DG 0.06 +/- 0.00
point	3 -	4,	DG 17.43 +/- 0.07	point	3 -	4,	DG 0.06 +/- 0.00
point	4 -	5,	DG 15.53 +/- 0.08	point	4 -	5,	DG 0.06 +/- 0.00
point	5 -	6,	DG 13.82 +/- 0.04	point	5 -	6,	DG 0.12 +/- 0.00
point	6 -	7,	DG 10.72 +/- 0.08	point	6 -	7,	DG 0.13 +/- 0.00
point	7 -	8,	DG 7.25 +/- 0.12	point	7 -	8,	DG 0.36 +/- 0.01
point	8 -	9,	DG 5.14 +/- 0.04	point	8 -	9,	DG 0.34 +/- 0.02
point	9 -	10,	DG 3.12 +/- 0.25	point	9 -	10,	DG 0.51 +/- 0.02
total	0 -	10,	DG 136.60 +/- 0.25	point	10 -	11,	DG 0.45 +/- 0.01
				total	0 -	11,	DG 2.17 +/- 0.05
Bound_coul = 32.64 kcal/mol				Bound_restr = 0.51 kcal/mol			
Complex = 35.44 kcal/mol							

Final results in kJ/mol:			
point	0 -	1,	DG 4.38 +/- 0.09
point	1 -	2,	DG 4.38 +/- 0.08
point	2 -	3,	DG 4.18 +/- 0.07
point	3 -	4,	DG 3.84 +/- 0.12
point	4 -	5,	DG 3.63 +/- 0.08
point	5 -	6,	DG 3.45 +/- 0.07
point	6 -	7,	DG 3.27 +/- 0.08
point	7 -	8,	DG 3.18 +/- 0.12
point	8 -	9,	DG 3.15 +/- 0.08
point	9 -	10,	DG 2.75 +/- 0.09
point	10 -	11,	DG 1.52 +/- 0.15
point	11 -	12,	DG 0.84 +/- 0.22
point	12 -	13,	DG 0.43 +/- 0.18
point	13 -	14,	DG -1.72 +/- 0.48
point	14 -	15,	DG -4.88 +/- 0.23
point	15 -	16,	DG -7.03 +/- 0.24
point	16 -	17,	DG -7.18 +/- 0.19
point	17 -	18,	DG -5.64 +/- 0.07
point	18 -	19,	DG -2.91 +/- 0.06
point	19 -	20,	DG -0.03 +/- 0.03
total	0 -	20,	DG 9.62 +/- 0.90

Bound_vdw = 2.29 kcal/mol

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 20.95 +/- 0.16	point	0 -	1,	DG 3.40 +/- 0.02
point	1 -	2,	DG 17.48 +/- 0.06	point	1 -	2,	DG 3.34 +/- 0.02
point	2 -	3,	DG 14.34 +/- 0.08	point	2 -	3,	DG 3.26 +/- 0.01
point	3 -	4,	DG 11.68 +/- 0.07	point	3 -	4,	DG 3.18 +/- 0.01
point	4 -	5,	DG 9.42 +/- 0.03	point	4 -	5,	DG 3.05 +/- 0.02
point	5 -	6,	DG 7.34 +/- 0.02	point	5 -	6,	DG 2.90 +/- 0.01
point	6 -	7,	DG 5.45 +/- 0.02	point	6 -	7,	DG 2.73 +/- 0.00
point	7 -	8,	DG 3.74 +/- 0.01	point	7 -	8,	DG 2.50 +/- 0.01
point	8 -	9,	DG 2.16 +/- 0.01	point	8 -	9,	DG 2.23 +/- 0.01
point	9 -	10,	DG 0.67 +/- 0.02	point	9 -	10,	DG 1.84 +/- 0.01
total	0 -	10,	DG 93.23 +/- 0.10	point	10 -	11,	DG 1.32 +/- 0.01
				point	11 -	12,	DG 0.62 +/- 0.03
				point	12 -	13,	DG -0.27 +/- 0.01
				point	13 -	14,	DG -1.67 +/- 0.01
				point	14 -	15,	DG -4.00 +/- 0.05
				point	15 -	16,	DG -7.40 +/- 0.08
				point	16 -	17,	DG -8.96 +/- 0.05
				point	17 -	18,	DG -7.32 +/- 0.02
				point	18 -	19,	DG -4.37 +/- 0.02
				point	19 -	20,	DG -1.39 +/- 0.01
				total	0 -	20,	DG -4.99 +/- 0.05

Unbound_coul= 22.28 kcal/mol

Unbound_vdw= -1.19 kcal/mol

Ligand in solvent = 21.09 kcal/mol

(ii) Run 3

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 23.36 +/- 0.04	point	0 -	1,	DG 0.03 +/- 0.00
point	1 -	2,	DG 21.11 +/- 0.06	point	1 -	2,	DG 0.04 +/- 0.00
point	2 -	3,	DG 19.10 +/- 0.11	point	2 -	3,	DG 0.07 +/- 0.00
point	3 -	4,	DG 17.32 +/- 0.10	point	3 -	4,	DG 0.07 +/- 0.00
point	4 -	5,	DG 14.64 +/- 0.36	point	4 -	5,	DG 0.06 +/- 0.00
point	5 -	6,	DG 11.41 +/- 0.28	point	5 -	6,	DG 0.13 +/- 0.00
point	6 -	7,	DG 9.01 +/- 0.11	point	6 -	7,	DG 0.12 +/- 0.00
point	7 -	8,	DG 6.75 +/- 0.16	point	7 -	8,	DG 0.34 +/- 0.00
point	8 -	9,	DG 4.88 +/- 0.17	point	8 -	9,	DG 0.32 +/- 0.00
point	9 -	10,	DG 3.93 +/- 0.12	point	9 -	10,	DG 0.48 +/- 0.00
total	0 -	10,	DG 131.52 +/- 1.21	point	10 -	11,	DG 0.47 +/- 0.01
				total	0 -	11,	DG 2.12 +/- 0.01
Bound_coul = 31.43 kcal/mol				Bound_restr = 0.5 kcal/mol			
Complex = 34.8 kcal/mol							

Final results in kJ/mol:			
point	0 -	1,	DG 4.34 +/- 0.05
point	1 -	2,	DG 4.43 +/- 0.06
point	2 -	3,	DG 4.28 +/- 0.06
point	3 -	4,	DG 4.15 +/- 0.08
point	4 -	5,	DG 4.17 +/- 0.07
point	5 -	6,	DG 3.81 +/- 0.12
point	6 -	7,	DG 3.72 +/- 0.09
point	7 -	8,	DG 3.83 +/- 0.10
point	8 -	9,	DG 3.28 +/- 0.08
point	9 -	10,	DG 2.46 +/- 0.09
point	10 -	11,	DG 2.20 +/- 0.10
point	11 -	12,	DG 1.38 +/- 0.15
point	12 -	13,	DG 0.21 +/- 0.21
point	13 -	14,	DG -1.36 +/- 0.29
point	14 -	15,	DG -4.97 +/- 0.32
point	15 -	16,	DG -7.56 +/- 0.12
point	16 -	17,	DG -7.63 +/- 0.14
point	17 -	18,	DG -5.76 +/- 0.05
point	18 -	19,	DG -2.91 +/- 0.03
point	19 -	20,	DG -0.05 +/- 0.07
total	0 -	20,	DG 12.03 +/- 1.35

Bound_vdw = 2.87 kcal/mol

Final results in kJ/mol:

point	0 -	1,	DG	20.98 +/-	0.10
point	1 -	2,	DG	17.35 +/-	0.06
point	2 -	3,	DG	14.33 +/-	0.05
point	3 -	4,	DG	11.70 +/-	0.03
point	4 -	5,	DG	9.41 +/-	0.02
point	5 -	6,	DG	7.33 +/-	0.01
point	6 -	7,	DG	5.42 +/-	0.01
point	7 -	8,	DG	3.73 +/-	0.01
point	8 -	9,	DG	2.18 +/-	0.01
point	9 -	10,	DG	0.69 +/-	0.01

total	0 -	10,	DG	93.12 +/-	0.12
-------	-----	-----	----	-----------	------

Unbound_coul = 22.25 kcal/mol

Ligand in solvent = 21 kcal/mol

Final results in kJ/mol:

point	0 -	1,	DG	3.40 +/-	0.01
point	1 -	2,	DG	3.35 +/-	0.01
point	2 -	3,	DG	3.27 +/-	0.00
point	3 -	4,	DG	3.17 +/-	0.01
point	4 -	5,	DG	3.04 +/-	0.01
point	5 -	6,	DG	2.89 +/-	0.01
point	6 -	7,	DG	2.71 +/-	0.00
point	7 -	8,	DG	2.50 +/-	0.00
point	8 -	9,	DG	2.22 +/-	0.01
point	9 -	10,	DG	1.83 +/-	0.01
point	10 -	11,	DG	1.35 +/-	0.01
point	11 -	12,	DG	0.65 +/-	0.02
point	12 -	13,	DG	-0.33 +/-	0.01
point	13 -	14,	DG	-1.71 +/-	0.02
point	14 -	15,	DG	-4.04 +/-	0.04
point	15 -	16,	DG	-7.50 +/-	0.02
point	16 -	17,	DG	-8.98 +/-	0.03
point	17 -	18,	DG	-7.29 +/-	0.01
point	18 -	19,	DG	-4.38 +/-	0.01
point	19 -	20,	DG	-1.40 +/-	0.00

total	0 -	20,	DG	-5.25 +/-	0.07
-------	-----	-----	----	-----------	------

Unbound_vdw = -1.25 kcal/mol

Appendix 6.7. Second and third ABFE runs on 5SOJ

(i) Run 2

				Final results in kJ/mol:													
Final results in kJ/mol:				point	0 -	1,	DG	4.92 +/-	0.06	Final results in kJ/mol:							
				point	1 -	2,	DG	4.85 +/-	0.05	point	0 -	1,	DG	0.03 +/-	0.00		
				point	2 -	3,	DG	4.90 +/-	0.06	point	1 -	2,	DG	0.05 +/-	0.00		
				point	3 -	4,	DG	4.71 +/-	0.02	point	2 -	3,	DG	0.08 +/-	0.00		
				point	4 -	5,	DG	4.41 +/-	0.02	point	3 -	4,	DG	0.08 +/-	0.00		
point	0 -	1,	DG	22.54 +/-	0.16	point	5 -	6,	DG	4.26 +/-	0.07	point	4 -	5,	DG	0.08 +/-	0.00
point	1 -	2,	DG	19.79 +/-	0.12	point	6 -	7,	DG	4.35 +/-	0.08	point	5 -	6,	DG	0.16 +/-	0.01
point	2 -	3,	DG	17.33 +/-	0.11	point	7 -	8,	DG	4.38 +/-	0.09	point	6 -	7,	DG	0.17 +/-	0.01
point	3 -	4,	DG	15.17 +/-	0.14	point	8 -	9,	DG	3.78 +/-	0.06	point	7 -	8,	DG	0.44 +/-	0.01
point	4 -	5,	DG	12.62 +/-	0.12	point	9 -	10,	DG	3.51 +/-	0.06	point	8 -	9,	DG	0.39 +/-	0.02
point	5 -	6,	DG	10.74 +/-	0.19	point	10 -	11,	DG	3.38 +/-	0.09	point	9 -	10,	DG	0.59 +/-	0.02
point	6 -	7,	DG	9.18 +/-	0.12	point	11 -	12,	DG	2.42 +/-	0.23	point	10 -	11,	DG	0.49 +/-	0.00
point	7 -	8,	DG	6.99 +/-	0.06	point	12 -	13,	DG	1.01 +/-	0.36	total	0 -	11,	DG	2.57 +/-	0.06
point	8 -	9,	DG	4.75 +/-	0.23	point	13 -	14,	DG	-0.96 +/-	0.42	Bound_restr = 0.61 kcal/mol					
point	9 -	10,	DG	3.12 +/-	0.24	point	14 -	15,	DG	-5.98 +/-	0.08						
				point	15 -	16,	DG	-8.97 +/-	0.21								
				point	16 -	17,	DG	-8.89 +/-	0.06								
				point	17 -	18,	DG	-6.70 +/-	0.06								
total	0 -	10,	DG	122.23 +/-	0.53	point	18 -	19,	DG	-3.36 +/-	0.07						
				point	19 -	20,	DG	-0.02 +/-	0.04								
				total	0 -	20,	DG	16.80 +/-	0.48								

Bound coul = 29.21 kcal/mol

Bound vdw = 4.01 kcal/mol

Complex = 33.83 kcal/mol

Final results in kJ/mol:

point	0 -	1,	DG	4.06 +/-	0.01
point	1 -	2,	DG	3.99 +/-	0.01
point	2 -	3,	DG	3.87 +/-	0.00
point	3 -	4,	DG	3.75 +/-	0.00
point	4 -	5,	DG	3.63 +/-	0.00
point	5 -	6,	DG	3.47 +/-	0.01
point	6 -	7,	DG	3.28 +/-	0.01
point	7 -	8,	DG	3.05 +/-	0.01
point	8 -	9,	DG	2.74 +/-	0.01
point	9 -	10,	DG	2.30 +/-	0.01
point	10 -	11,	DG	1.77 +/-	0.01
point	11 -	12,	DG	1.06 +/-	0.01
point	12 -	13,	DG	0.07 +/-	0.02
point	13 -	14,	DG	-1.50 +/-	0.04
point	14 -	15,	DG	-4.34 +/-	0.04
point	15 -	16,	DG	-8.85 +/-	0.09
point	16 -	17,	DG	-10.65 +/-	0.05
point	17 -	18,	DG	-8.51 +/-	0.02
point	18 -	19,	DG	-5.08 +/-	0.01
point	19 -	20,	DG	-1.63 +/-	0.01
total	0 -	20,	DG	-3.52 +/-	0.17

Unbound coul = 24.97 kcal/mol

Ligand in solvent = 24.13 kcal/mol Unbound vdw = -0.84 kcal/mol

(ii) Run 3

```

Final results in kJ/mol:
point 0 - 1, DG 4.99 +/- 0.04
point 1 - 2, DG 4.96 +/- 0.07
point 2 - 3, DG 4.78 +/- 0.02
point 3 - 4, DG 4.59 +/- 0.06
point 4 - 5, DG 4.49 +/- 0.02
point 5 - 6, DG 4.51 +/- 0.07
point 6 - 7, DG 4.26 +/- 0.06
point 7 - 8, DG 3.86 +/- 0.06
point 8 - 9, DG 3.58 +/- 0.10
point 9 - 10, DG 3.51 +/- 0.07
point 10 - 11, DG 3.23 +/- 0.07
point 11 - 12, DG 2.67 +/- 0.09
point 12 - 13, DG 1.86 +/- 0.13
point 13 - 14, DG -1.10 +/- 0.38
point 14 - 15, DG -5.77 +/- 0.43
point 15 - 16, DG -8.89 +/- 0.18
point 16 - 17, DG -9.18 +/- 0.18
point 17 - 18, DG -6.77 +/- 0.07
point 18 - 19, DG -3.36 +/- 0.02
point 19 - 20, DG 0.04 +/- 0.03
total 0 - 20, DG 16.33 +/- 1.10

Final results in kJ/mol:
point 0 - 1, DG 0.04 +/- 0.00
point 1 - 2, DG 0.05 +/- 0.00
point 2 - 3, DG 0.08 +/- 0.01
point 3 - 4, DG 0.09 +/- 0.01
point 4 - 5, DG 0.08 +/- 0.01
point 5 - 6, DG 0.14 +/- 0.00
point 6 - 7, DG 0.15 +/- 0.00
point 7 - 8, DG 0.43 +/- 0.01
point 8 - 9, DG 0.39 +/- 0.01
point 9 - 10, DG 0.62 +/- 0.03
point 10 - 11, DG 0.55 +/- 0.01
total 0 - 11, DG 2.63 +/- 0.06

Bound_restr = 0.62 kcal/mol

```

Bound_coul = 28.78 kcal/mol

Bound_vdw = 3.9 kcal/mol

Complex = 33.3 kcal/mol

Final results in kJ/mol:

point	0 -	1,	DG	22.63 +/-	0.02
point	1 -	2,	DG	19.40 +/-	0.02
point	2 -	3,	DG	16.35 +/-	0.02
point	3 -	4,	DG	13.49 +/-	0.03
point	4 -	5,	DG	10.79 +/-	0.02
point	5 -	6,	DG	8.33 +/-	0.01
point	6 -	7,	DG	6.14 +/-	0.02
point	7 -	8,	DG	4.17 +/-	0.02
point	8 -	9,	DG	2.40 +/-	0.01
point	9 -	10,	DG	0.75 +/-	0.01
total	0 -	10,	DG	104.44 +/-	0.10

Unbound_coul = 24.96 kcal/mol

Ligand in solvent = 24.16 kcal/mol

Final results in kJ/mol:

point	0 -	1,	DG	4.04 +/-	0.01
point	1 -	2,	DG	3.97 +/-	0.01
point	2 -	3,	DG	3.87 +/-	0.00
point	3 -	4,	DG	3.76 +/-	0.01
point	4 -	5,	DG	3.63 +/-	0.01
point	5 -	6,	DG	3.47 +/-	0.01
point	6 -	7,	DG	3.28 +/-	0.01
point	7 -	8,	DG	3.04 +/-	0.00
point	8 -	9,	DG	2.75 +/-	0.01
point	9 -	10,	DG	2.34 +/-	0.01
point	10 -	11,	DG	1.79 +/-	0.01
point	11 -	12,	DG	1.10 +/-	0.01
point	12 -	13,	DG	0.07 +/-	0.03
point	13 -	14,	DG	-1.54 +/-	0.04
point	14 -	15,	DG	-4.34 +/-	0.04
point	15 -	16,	DG	-8.79 +/-	0.10
point	16 -	17,	DG	-10.66 +/-	0.06
point	17 -	18,	DG	-8.48 +/-	0.03
point	18 -	19,	DG	-5.06 +/-	0.02
point	19 -	20,	DG	-1.61 +/-	0.02
total	0 -	20,	DG	-3.37 +/-	0.25

Unbound_vdw = -0.80 kcal/mol

Appendix 6.8. Second and third ABFE runs on CACHE_1689_35

(i) Run 2

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 18.01 +/- 0.05	point	0 -	1,	DG 0.27 +/- 0.06
point	1 -	2,	DG 15.63 +/- 0.07	point	1 -	2,	DG 0.06 +/- 0.00
point	2 -	3,	DG 13.24 +/- 0.07	point	2 -	3,	DG 0.09 +/- 0.00
point	3 -	4,	DG 11.14 +/- 0.03	point	3 -	4,	DG 0.12 +/- 0.03
point	4 -	5,	DG 9.31 +/- 0.03	point	4 -	5,	DG 0.12 +/- 0.02
point	5 -	6,	DG 7.38 +/- 0.07	point	5 -	6,	DG 0.18 +/- 0.01
point	6 -	7,	DG 5.63 +/- 0.10	point	6 -	7,	DG 0.16 +/- 0.01
point	7 -	8,	DG 4.28 +/- 0.06	point	7 -	8,	DG 0.42 +/- 0.01
point	8 -	9,	DG 3.08 +/- 0.03	point	8 -	9,	DG 0.39 +/- 0.00
point	9 -	10,	DG 1.99 +/- 0.05	point	9 -	10,	DG 0.61 +/- 0.02
total	0 -	10,	DG 89.70 +/- 0.29	point	10 -	11,	DG 0.56 +/- 0.02
Bound_coul = 21.43 kcal/mol				total	0 -	11,	DG 2.97 +/- 0.09
Complex = 27.43 kcal/mol				Bound_restr = 0.7 kcal/mol			
				total	0 -	20,	DG 22.22 +/- 1.13
				Bound_vdw = 5.3 kcal/mol			

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 15.98 +/- 0.06	point	0 -	1,	DG 4.54 +/- 0.02
point	1 -	2,	DG 13.53 +/- 0.08	point	1 -	2,	DG 4.42 +/- 0.02
point	2 -	3,	DG 11.20 +/- 0.08	point	2 -	3,	DG 4.29 +/- 0.02
point	3 -	4,	DG 9.08 +/- 0.05	point	3 -	4,	DG 4.16 +/- 0.03
point	4 -	5,	DG 7.24 +/- 0.06	point	4 -	5,	DG 4.03 +/- 0.03
point	5 -	6,	DG 5.56 +/- 0.04	point	5 -	6,	DG 3.89 +/- 0.03
point	6 -	7,	DG 4.04 +/- 0.02	point	6 -	7,	DG 3.69 +/- 0.02
point	7 -	8,	DG 2.82 +/- 0.04	point	7 -	8,	DG 3.43 +/- 0.01
point	8 -	9,	DG 1.63 +/- 0.04	point	8 -	9,	DG 3.07 +/- 0.02
point	9 -	10,	DG 0.49 +/- 0.03	point	9 -	10,	DG 2.67 +/- 0.00
total	0 -	10,	DG 71.56 +/- 0.18	point	10 -	11,	DG 2.11 +/- 0.03
Unbound_coul = 17.10 kcal/mol				point	11 -	12,	DG 1.30 +/- 0.02
				point	12 -	13,	DG 0.17 +/- 0.02
				point	13 -	14,	DG -1.62 +/- 0.04
				point	14 -	15,	DG -4.87 +/- 0.05
				point	15 -	16,	DG -9.40 +/- 0.03
				point	16 -	17,	DG -11.12 +/- 0.03
				point	17 -	18,	DG -8.88 +/- 0.02
				point	18 -	19,	DG -5.29 +/- 0.01
				point	19 -	20,	DG -1.71 +/- 0.01
				total	0 -	20,	DG -1.11 +/- 0.16
				Ligand in solvent = 16.84 kcal/mol			
				Unbound_vdw = -0.26 kcal/mol			

(ii) Run 3

Final results in kJ/mol:				Final results in kJ/mol:			
point	0 -	1,	DG 17.88 +/- 0.13	point	0 -	1,	DG 0.58 +/- 0.07
point	1 -	2,	DG 15.66 +/- 0.12	point	1 -	2,	DG 1.04 +/- 0.25
point	2 -	3,	DG 13.29 +/- 0.08	point	2 -	3,	DG 0.10 +/- 0.01
point	3 -	4,	DG 11.20 +/- 0.17	point	3 -	4,	DG 0.09 +/- 0.00
point	4 -	5,	DG 9.28 +/- 0.09	point	4 -	5,	DG 0.09 +/- 0.00
point	5 -	6,	DG 7.40 +/- 0.02	point	5 -	6,	DG 0.17 +/- 0.00
point	6 -	7,	DG 5.66 +/- 0.04	point	6 -	7,	DG 0.17 +/- 0.01
point	7 -	8,	DG 4.40 +/- 0.04	point	7 -	8,	DG 0.49 +/- 0.02
point	8 -	9,	DG 3.13 +/- 0.03	point	8 -	9,	DG 0.40 +/- 0.01
point	9 -	10,	DG 1.71 +/- 0.02	point	9 -	10,	DG 0.57 +/- 0.01
total	0 -	10,	DG 89.62 +/- 0.25	point	10 -	11,	DG 0.51 +/- 0.01
Bound_coul = 21.41 kcal/mol				total	0 -	11,	DG 4.20 +/- 0.31
				Bound_restr = 1.003 kcal/mol			
				total	0 -	20,	DG 23.16 +/- 0.87
				Bound_vdw = 5.5 kcal/mol			
				Complex = 27.91 kcal/mol			

Final results in kJ/mol:

point	0 -	1,	DG	16.09 +/-	0.07
point	1 -	2,	DG	13.42 +/-	0.05
point	2 -	3,	DG	11.11 +/-	0.08
point	3 -	4,	DG	9.12 +/-	0.08
point	4 -	5,	DG	7.18 +/-	0.07
point	5 -	6,	DG	5.43 +/-	0.05
point	6 -	7,	DG	4.06 +/-	0.03
point	7 -	8,	DG	2.77 +/-	0.03
point	8 -	9,	DG	1.56 +/-	0.04
point	9 -	10,	DG	0.61 +/-	0.05
total	0 -	10,	DG	71.35 +/-	0.44

Unbound_coul = 17.05 kcal/mol

Ligand in solvent = **16.88 kcal/mol** Unbound_vdw = -0.17 kcal/mol

Final results in kJ/mol:

point	0 -	1,	DG	4.53 +/-	0.04
point	1 -	2,	DG	4.44 +/-	0.03
point	2 -	3,	DG	4.31 +/-	0.01
point	3 -	4,	DG	4.20 +/-	0.02
point	4 -	5,	DG	4.08 +/-	0.01
point	5 -	6,	DG	3.86 +/-	0.01
point	6 -	7,	DG	3.67 +/-	0.02
point	7 -	8,	DG	3.45 +/-	0.02
point	8 -	9,	DG	3.12 +/-	0.01
point	9 -	10,	DG	2.66 +/-	0.02
point	10 -	11,	DG	2.10 +/-	0.01
point	11 -	12,	DG	1.31 +/-	0.03
point	12 -	13,	DG	0.23 +/-	0.03
point	13 -	14,	DG	-1.51 +/-	0.07
point	14 -	15,	DG	-4.75 +/-	0.09
point	15 -	16,	DG	-9.40 +/-	0.04
point	16 -	17,	DG	-11.08 +/-	0.03
point	17 -	18,	DG	-8.90 +/-	0.04
point	18 -	19,	DG	-5.33 +/-	0.02
point	19 -	20,	DG	-1.71 +/-	0.00
total	0 -	20,	DG	-0.72 +/-	0.19

Appendix 6.9.

Publications (Peer-reviewed)

1. Malebari A M, **Kandwal S**, Ali A, Fayne D, Twamley B, Zisterer D M, & Meegan M J. Synthesis, Computational Studies, and Structural Analysis of 1-(3,5-Dimethoxyphenyl)azetidin-2-ones with Antiproliferative Activity in Breast Cancer and Chemoresistant Colon Cancer. *Pharmaceuticals*. **2025**, 18(9), 1330.
2. Herasymenko O, Silva M, Abu-Saleh AAA, Ahmad A, Alvarado-Huayhuaz J, Arce OEA, Armstrong RJ, Arrowsmith C, Bachta KE, Beck H, Berta D, Bieniek MK, Blay V, Bolotokova A, Bourne PE, Breznik M, Brown PJ, Campbell ADG, Carosati E, Chau I, Cole DJ, Cree B, Dehaen W, Denzinger K, Dos Santos Machado K, Dunn I, Durai P, Edfeldt K, Edwards A, Fayne D, Felfoldi D, Friston K, Ghiabi P, Gibson E, Günther J, Gunnarsson A, Hillisch A, Houston DR, Jensen JH, Harding RJ, Harris KS, Hoffer L, Hogner A, Horton JT, Houliston S, Hultquist JF, Hutchinson A, Irwin JJ, Jukič M, **Kandwal S**, Karlova A, Katis VL, Kich RP, Kireev D, Koes D, Inniss NL, Lessel U, Liu S, Loppnau P, Lu W, Martino S, McGibbon M, Meiler J, Mettu A, Money-Kyrle S, Moretti R, Moroz YS, Muvva C, Newman JA, Obendorf L, Paige B, Pandit A, Park K, Perveen S, Pirie R, Poda G, Protopopov M, Pütter V, Ricci F, Roper NJ, Rosta E, Rzhetskaya M, Sabnis Y, Satchell KJF, Schmitt Kremer F, Scott T, Seitova A, Steinmann C, Talagayev V, Tarkhanova OO, Tatum NJ, Treleaven D, Velasque Werhli A, Walters WP, Wang X, Wells J, Wells G, Westermaier Y, Wolber G, Wortmann L, Zhang J, Zhao Z, Zheng S, Schapira M. CACHE Challenge #2: Targeting the RNA Site of the SARS-CoV-2 Helicase Nsp13. *J Chem Inf Model*. **2025**, 14;65(13):6884-6898
3. Zhang Y, **Kandwal S**, Fayne D, Stevenson NJ. MERS-CoV-nsp5 expression in human epithelial BEAS 2b cells attenuates type I interferon production by inhibiting IRF3 nuclear translocation. *Cell Mol Life Sci*. **2024**, 81(1), 433
4. McKeown J P, Byrne A J, Bright S A, Charleton, C E, **Kandwal S**, Čmelo I, Twamley B, McElligott A M, Fayne D, O'Boyle N M, et al. Synthesis and Biochemical Evaluation of Ethanoanthracenes and Related Compounds: Antiproliferative and Pro-Apoptotic Effects in Chronic Lymphocytic Leukemia (CLL). *Pharmaceuticals* **2024**, 17, 1034
5. Efstathiou C, Zhang Y, **Kandwal S**, Fayne D, Molloy EJ, Stevenson NJ. Respiratory syncytial virus NS1 inhibits anti-viral Interferon- α -induced JAK/STAT signaling, by limiting the nuclear translocation of STAT1. *Front Immunol*. **2024**, 15-1395809
6. Byrne AJ, Bright SA, McKeown JP, Bergin A, Twamley B, McElligott AM, Noorani S, **Kandwal S**, Fayne D, O'Boyle NM, Williams DC, Meegan MJ. Synthesis and Pro-Apoptotic Effects of Nitrovinylanthracenes and Related Compounds in Chronic Lymphocytic Leukaemia (CLL) and Burkitt's Lymphoma (BL). *Molecules*. **2023** 28(24), 8095

7. Khan MF^a, **Kandwal S^a**, Fayne D. DataPype: A Fully Automated Unified Software Platform for Computer-Aided Drug Design. ACS Omega. **2023** 8(42), 39468-39480 ^a**Joint first authors**
8. Wang S, Malebari AM, Greene TF, **Kandwal S**, Fayne D, Nathwani SM, Zisterer DM, Twamley B, O'Boyle NM, Meegan MJ. Antiproliferative and Tubulin-Destabilising Effects of 3-(Prop-1-en-2-yl)azetidin-2-Ones and Related Compounds in MCF-7 and MDA-MB-231 Breast Cancer Cells. Pharmaceuticals (Basel). **2023** 16(7), 1000
9. **Kandwal S**, Fayne D. Genetic conservation across SARS-CoV-2 non-structural proteins - Insights into possible targets for treatment of future viral outbreaks. Virology. **2023** 581, 97-115

Pre-print (submitted for peer-review)

1. **Kandwal, S.**, Cmelo, I., & Fayne, D. SARS-CoV-2 ORF8 sequence conservation and mutational analysis—Insight into the influence of dataset size on identifying top mutations. *bioRxiv*, **2025**.06.16.659836
2. Herasymenko, O., Silva, M., Correy, G.J., Abu-Saleh, A.A.-A.A., Ackloo, S., Arrowsmith, C., Ban, F., Beck, H., Bishop, K.P., Bohórquez, H.J., Bolotokova, A., Breznik, M., Chau, I., Chen, Y., Cherkasov, A., Dehaen, W., Corte, D.D., Denzinger, K., Edfeldt, K., Edwards, A., Fayne, D., Gentile, F., Gibson, E., Gokdemir, O., Irwin, J.J., Gunnarsson, A., Günther, J., Jensen, J.H., Harding, R.J., Hillisch, A., Hoffer, L., Hogner, A., Hutchinson, A., **Kandwal, S.**, Karlova, A., Kozakov, D., Lee, J., Lee, S., Lessel, U., Liu, S., Liu, X., Loppnau, P., Meiler, J., Moretti, R., Moroz, Y.S., Muvva, C., Oprea, T.I., Paige, B., Pandit, A., Park, K., Poda, G., Protopopov, M.V., Pütter, V., Ravichandran, R., Rognan, D., Rosta, E., Sabnis, Y., Scott, T., Seitova, A., Sharma, P., Sindt, F., Song, M., Steinmann, C., Stevens, R., Talagayev, V., Tararina, V.V., Tarkhanova, O., Tingey, D., Trant, J., Treleaven, D., Tropsha, A., Walters, P., Wells, J., Westermaier, Y., Wolber, G., Wortmann, L., Zheng, S., Fraser, J.S., Schapira, M. CACHE Challenge #3: Targeting the Nsp3 Macrodomein of SARS-CoV-2. ChemRxiv. **2025**

Presentations

- **Oral presentation** at 75th Irish Universities Chemistry Research Colloquium on 17th – 18th June 2024. Title: Discovery of small molecules blocking a key binding site in SARS-CoV-2 NSP3 protein.
- **Oral presentation** at 12th meeting of the Dublin academy of pathogenomics and infection biology on 07th June 2024. Title: Discovery of small molecules blocking a key binding site in SARS-CoV-2 NSP3 protein.
- **Oral presentation** at 9th Chemistry Day, Dublin City University on 10th May 2024.

- **Flash talk 3 slides in 3 mins:** Presented my research at 2023 cohort Tipton medal competition on 24th April 2024.
- **Oral presentation** (invited speaker) at the International Pandemic Sciences Conference, University of Oxford on the 11th of July 2023. Title: Sequence and pocket conservation across SARS-CoV-2 non-structural proteins - design of future therapeutics (NSP13).
- **Oral presentation** (invited speaker) at the Biochemical Society's Early Career Researchers (ECR) webinar series 'Drug discovery and structure-function relationships' on 24th of April 2023. Title: Sequence and pocket conservation across SARS-CoV-2 non-structural proteins - design of future therapeutics (NSP3). Link: <https://youtu.be/oEL75hrAWjg?si=7QPBRItCDoybVVJQ>
- **Poster presentation with 2 mins flash talk** at 28th MGMS Young Modellers' Forum 2024 on 29th November 2024.
- **Poster presentations:** Presented posters at the following conferences: -

Title - Bioinformatics and cheminformatics analysis on SARS-CoV-2 sequences to identify conserved binding pockets across the viral non-structural proteins - design of future therapeutics.

1. Irish Association of Pharmacologists (IAP) Annual Meeting - 11/11/2022
2. CECAM Mixed-Gen Season 3 - Session 5: Data science in biophysics for applications to drug discovery - 23/03/2023

Outreach Activities

- Research on walls competition, Dublin City University on 11th October 2024 - Captured my research in one picture.
- Research visualization competition, Dublin City University on 17th April 2024 - PhD research in graphical abstract format.
- IRC 2023 video competition: Why research matters?
Link: <https://youtu.be/lkNfm8ejGH8?si=bEih5vRadruCULmW>
- HERStory Makers 2023: Participated in the HERStory Makers competition for female-identifying early career researchers to share their research on 28th September 2023. Link: <https://doi.org/10.6084/m9.figshare.24158688.v1>
- START European researchers' night: Presented a poster describing my research project 27/09/2023-28/09/2023.

Recognition

- Best poster presenter with 2 min flash talk at the 28th MGMS Young Modellers' Forum (YMF) 2024 at University of Oxford.

- Secured the Notre Dame-DCU Inaugural Seed Grant in Precision Biomedical Technologies 2024.
- CACHE Challenges - Our group was ranked among the top 4 globally to computationally identify small molecules that bind to the key binding sites of SARS-CoV-2 NSP3 and NSP13.
- Runner up in long presentation category at 9th Chemistry Day at Dublin City University.
- Awarded bursary to attend and present my research at the International Pandemic Sciences Conference, University of Oxford.

Courses completed

- Understanding and Visualizing Data with python - University of Michigan (Coursera)
- Learn Bioinformatics from Scratch (Theory & Practical) – Udemy
- Research Project Management – SciLink
- Bioinformatics tools for research – Udemy
- Python for beginners: Learn all the basics of python – Udemy
- Research Integrity and Impact in an Open Scholarship Era Module **(5 ECTS)** – TCD
- Teaching and Supporting Learning as a Graduate Teach Assistant Module **(5 ECTS)** - TCD

References

- Abdelnabi, R., Neyts, J., & Delang, L. (2015). Towards antivirals against chikungunya virus. *Antiviral Research*, 121, 59–68. <https://doi.org/10.1016/j.antiviral.2015.06.017>
- Adedeji A.O., Singh K., Calcaterra N.E., DeDiego M.L., Enjuanes L., Weiss S., Sarafianos S.G. Severe acute respiratory syndrome coronavirus replication inhibitor that interferes with the nucleic acid unwinding of the viral helicase. *Antimicrob. Agents Chemother.* 2012;56(9):4718–4728. doi: 10.1128/AAC.00957-12.
- Afsar M, Narayan R, Akhtar MN, Das D, Rahil H, Nagaraj SK, Eswarappa SM, Tripathi S, Hussain T. Drug targeting Nsp1-ribosomal complex shows antiviral activity against SARS-CoV-2. *Elife*. 2022 Mar 24;11:e74877. doi: 10.7554/eLife.74877. PMID: 35323109; PMCID: PMC9018067.
- Agu, P. C., Afiukwa, C. A., Orji, O. U., Ezeh, E. M., Ofoke, I. H., Ogbu, C. O., Ugwuja, E. I., & Aja, P. M. (2023). Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. *Scientific reports*, 13(1), 13398. <https://doi.org/10.1038/s41598-023-40160-2>
- Ahola, T., & Kääriäinen, L. (1995). Reaction in alphavirus mRNA capping: Formation of a covalent complex of nonstructural protein nsP1 with 7-methyl-GMP. *Proceedings of the National Academy of Sciences*, 92(2), 507–511. <https://doi.org/10.1073/pnas.92.2.507>
- Aldeghi M, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem Sci*. 2016 Jan 14;7(1):207-218. doi: 10.1039/c5sc02678d. Epub 2015 Oct 7. PMID: 26798447; PMCID: PMC4700411.
- Aldeghi, M., Bluck, J. P., & Biggin, P. C. (2018). Absolute Alchemical Free Energy Calculations for Ligand Binding: A Beginner's Guide. *Methods in molecular biology (Clifton, N.J.)*, 1762, 199–232. https://doi.org/10.1007/978-1-4939-7756-7_11
- Alhammad, Y. M. O., Kashipathy, M. M., Roy, A., Gagné, J.-P., McDonald, P., Gao, P., Nonfoux, L., Battaile, K. P., Johnson, D. K., Holmstrom, E. D., Poirier, G. G., Lovell, S., & Fehr, A. R. (2021). The SARS-CoV-2 Conserved Macrodome Is a Mono-ADP-Ribosylhydrolase. *Journal of Virology*, 95(3), e01969-20. <https://doi.org/10.1128/JVI.01969-20>
- Alhammad YM, Parthasarathy S, Ghimire R, O'Connor JJ, Kerr CM, Pfannenstiel JJ, Chanda D, Miller CA, Unckless RL, Zuniga S, Enjuanes L, More S, Channappanavar R, Fehr AR. SARS-CoV-2 Mac1 is required for IFN antagonism and efficient virus replication in mice. *bioRxiv [Preprint]*. 2023 Apr 6:2023.04.06.535927. doi: 10.1101/2023.04.06.535927. Update in: *Proc Natl Acad Sci U S A*. 2023 Aug 29;120(35):e2302083120. doi: 10.1073/pnas.2302083120. PMID: 37066301; PMCID: PMC10104158.
- Alibay I, Magarkar A, Seeliger D, Biggin PC. Evaluating the use of absolute binding free energy in the fragment optimisation process. *Commun Chem*. 2022 Sep 5;5(1):105. doi: 10.1038/s42004-022-00721-4. PMID: 36697714; PMCID: PMC9814858.
- Almeida M.S., Johnson M.A., Herrmann T., Geralt M., Wüthrich K. Novel beta-barrel fold in the nuclear magnetic resonance structure of the replicase nonstructural protein 1 from the severe acute respiratory syndrome coronavirus. *J. Virol.* 2007;81(7):3151–3161. doi: 10.1128/JVI.01939-06.
- Almubaid Z., Al-Mubaid H. Analysis and comparison of genetic variants and mutations of the novel coronavirus SARS-CoV-2. *Gene Rep.* 2021;23 doi: 10.1016/j.genrep.2021.101064.

- Anand K., Palm G.J., Mesters J.R., Siddell S.G., Ziebuhr J., Hilgenfeld R. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J.* 2002;21(13):3213–3224. doi: 10.1093/emboj/cdf327.
- Anand K., Ziebuhr J., Wadhwani P., Mesters J.R., Hilgenfeld R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science (New York, N.Y.)* 2003;300(5626):1763–1767. doi: 10.1126/science.1085658.
- Báez-Santos Y.M., St John S.E., Mesecar A.D. The SARS-coronavirus papain-like protease: structure, function and inhibition by designed antiviral compounds. *Antivir. Res.* 2015;115:21–38. doi: 10.1016/j.antiviral.2014.12.015.
- Bafna K., White K., Harish B., Rosales R., Ramelot T.A., Acton T.B., Moreno E., Kehrer T., Miorin L., Royer C.A., García-Sastre A., Krug R.M., Montelione G.T. Hepatitis C virus drugs that inhibit SARS-CoV-2 papain-like protease synergize with remdesivir to suppress viral replication in cell culture. *Cell Rep.* 2021;35(7) doi: 10.1016/j.celrep.2021.109133.
- Bafna Khushboo, et al. Structural similarity of SARS-CoV2 m^{pro} and HCV NS3/4A proteases suggests new approaches for identifying existing drugs useful as COVID-19 therapeutics. *Chem.* 2020 doi: 10.26434/chemrxiv.12153615. the preprint server for chemistry 10.26434/chemrxiv.12153615.v1. (Preprint)
- Baig, M. H., Ahmad, K., Rabbani, G., Danishuddin, M., & Choi, I. (2018). Computer Aided Drug Design and its Application to the Development of Potential Drugs for Neurodegenerative Disorders. *Current neuropharmacology*, 16(6), 740–748. <https://doi.org/10.2174/1570159X15666171016163510>
- Baltimore D. Expression of animal virus genomes. *Bacteriol. Rev.* 1971;35(3):235–241. doi: 10.1128/br.35.3.235-241.1971.
- Bartholomeeusen, K., Daniel, M., LaBeaud, D. A., Gasque, P., Peeling, R. W., Stephenson, K. E., Ng, L. F. P., & Ariën, K. K. (2023). Chikungunya fever. *Nature Reviews Disease Primers*, 9(1), 17. <https://doi.org/10.1038/s41572-023-00429-2>
- Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1), 164–171. <http://www.jstor.org/stable/2239727>
- Bayat A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ (Clinical research ed.)*, 324(7344), 1018–1022. <https://doi.org/10.1136/bmj.324.7344.1018>
- Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8), 3684–3690. <https://doi.org/10.1063/1.448118>
- Berman J.J. Group IV viruses: single-stranded (+)Sense RNA. *Taxonomic. Guide Infect Dis.* 2012:237–246. doi: 10.1016/B978-0-12-415895-5.00042-8.
- Biswal M., Diggs S., Xu D., Khudaverdyan N., Lu J., Fang J., Blaha G., Hai R., Song J. Two conserved oligomer interfaces of NSP7 and NSP8 underpin the dynamic assembly of SARS-CoV-2 RdRP. *Nucleic Acids Res.* 2021;49(10):5956–5966. doi: 10.1093/nar/gkab370.
- Blackshields G, Sievers F, Shi W, Wilm A, Higgins DG. Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms Mol Biol.* 2010 May 14;5:21. doi: 10.1186/1748-7188-5-21. PMID: 20470396; PMCID: PMC2893182.

- Bobrovs R., Kanepe I., Narvaiss N., Patetko L., Kalnins G., Sisovs M., Bula A.L., Grinberga S., Boroduskis M., Ramata-Stunda A., Rostoks N., Jirgensons A., Tars K., Jaudzems K. Discovery of SARS-CoV-2 Nsp14 and Nsp16 methyltransferase inhibitors by high-throughput virtual screening. *Pharmaceuticals*. 2021;14(12):1243. doi: 10.3390/ph14121243.
- Boresch, S., Tettinger, F., Leitgeb, M., & Karplus, M. (2003). Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *The Journal of Physical Chemistry B*, 107(35), 9535–9551. <https://doi.org/10.1021/jp0217839>
- Bouvet M., Debarnot C., Imbert I., Selisko B., Snijder E.J., Canard B., Decroly E. In vitro reconstitution of SARS-coronavirus mRNA cap methylation. *PLoS Pathog.* 2010;6(4) doi: 10.1371/journal.ppat.1000863.
- Bouvet M., Imbert I., Subissi L., Gluais L., Canard B., Decroly E. RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. *Proc. Natl. Acad. Sci. U.S.A.* 2012;109(24):9372–9377. doi: 10.1073/pnas.1201130109.
- Braun, J., & Fayne, D. (2022). Mapping of Protein Binding Sites using clustering algorithms - Development of a pharmacophore based drug discovery tool. *Journal of molecular graphics & modelling*, 115, 108228. <https://doi.org/10.1016/j.jmgm.2022.108228>
- Briolant, S., Garin, D., Scaramozzino, N., Jouan, A., & Crance, J. M. (2004). In vitro inhibition of Chikungunya and Semliki Forest viruses replication by antiviral compounds: Synergistic effect of interferon- α and ribavirin combination. *Antiviral Research*, 61(2), 111–117. <https://doi.org/10.1016/j.antiviral.2003.09.005>
- Brooks, B. R., Brooks, C. L., MacKerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., ... Karplus, M. (2009). CHARMM: The Biomolecular Simulation Program. *Journal of Computational Chemistry*, 30(10), 1545–1614. <https://doi.org/10.1002/jcc.21287>
- Brown, R. S., Anastasakis, D. G., Hafner, M., & Kielian, M. (2020). Multiple capsid protein binding sites mediate selective packaging of the alphavirus genomic RNA. *Nature Communications*, 11(1), 4693. <https://doi.org/10.1038/s41467-020-18447-z>
- Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1), 014101. <https://doi.org/10.1063/1.2408420>
- Cai Q., Yang M., Liu D., Chen J., Shu D., Xia J., Liao X., Gu Y., Cai Q., Yang Y., Shen C., Li X., Peng L., Huang D., Zhang J., Zhang S., Wang F., Liu J., Chen L., Chen S., et al. Experimental treatment with favipiravir for COVID-19: an open-label control study. *Engineering*. 2020;6(10):1192–1198. doi: 10.1016/j.eng.2020.03.007.
- Case, D. A., Cheatham III, T. E., Darden, T., Gohlke, H., Luo, R., Merz Jr., K. M., Onufriev, A., Simmerling, C., Wang, B., & Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16), 1668–1688. <https://doi.org/10.1002/jcc.20290>
- Ceraolo C., Giorgi F.M. Genomic variance of the 2019-nCoV coronavirus. *J. Med. Virol.* 2020;92(5):522–528. doi: 10.1002/jmv.25700.
- Chan A.P., Choi Y., Schork N.J. Conserved genomic terminals of SARS-CoV-2 as Co-evolving functional elements and potential therapeutic targets. *bioRxiv : Preprint Server Biol.* 2020;2020doi: 10.1101/2020.07.06.190207. 07.06.190207.

- Chan J.F., Kok K.H., Zhu Z., Chu H., To K.K., Yuan S., Yuen K.Y. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microb. Infect.* 2020;9(1):221–236. doi: 10.1080/22221751.2020.1719902.
- Chen J., Malone B., Llewellyn E., Grasso M., Shelton P., Olinares P., Maruthi K., Eng E.T., Vatandaslar H., Chait B.T., Kapoor T.M., Darst S.A., Campbell E.A. Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex. *Cell*. 2020;182(6):1560–1573. doi: 10.1016/j.cell.2020.07.033. e13.
- Chen S.C., Olsthoorn R.C. Group-specific structural features of the 5'-proximal sequences of coronavirus genomic RNAs. *Virology*. 2010;401(1):29–41. doi: 10.1016/j.virol.2010.02.007.
- Chen Z., Boon S.S., Wang M.H., Chan R., Chan P. Genomic and evolutionary comparison between SARS-CoV-2 and other human coronaviruses. *J. Virol Methods*. 2021;289doi: 10.1016/j.jviromet.2020.114032.
- Chen, R., Mukhopadhyay, S., Merits, A., Bolling, B., Nasar, F., Coffey, L. L., Powers, A., Weaver, S. C., & ICTV Report Consortium. (2018). ICTV Virus Taxonomy Profile: Togaviridae. *Journal of General Virology*, 99(6), 761–762. <https://doi.org/10.1099/jgv.0.001072>
- Cheng, R. H., Kuhn, R. J., Olson, N. H., Rossmann, M. G., Choi, H.-K., Smith, T. J., & Baker, T. S. (1995). Nucleocapsid and Glycoprotein Organization in an Enveloped Virus. *Cell*, 80(4), 621–630. [https://doi.org/10.1016/0092-8674\(95\)90516-2](https://doi.org/10.1016/0092-8674(95)90516-2)
- Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P. M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., Cofer, E. M., ... Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface*, 15(141), 20170387. <https://doi.org/10.1098/rsif.2017.0387>
- Chopra, A., Saluja, M., & Venugopalan, A. (2014). Effectiveness of Chloroquine and Inflammatory Cytokine Response in Patients With Early Persistent Musculoskeletal Pain and Arthritis Following Chikungunya Virus Infection. *Arthritis & Rheumatology*, 66(2), 319–326. <https://doi.org/10.1002/art.38221>
- Chou C.Y., Chien C.H., Han Y.S., Prebanda M.T., Hsieh H.P., Turk B., Chang G.G., Chen X. Thiopurine analogues inhibit papain-like protease of severe acute respiratory syndrome coronavirus. *Biochem. Pharmacol.* 2008;75(8):1601–1609. doi: 10.1016/j.bcp.2008.01.005.
- Chowdhury, B., & Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 109(5), 419–431. <https://doi.org/10.1016/j.ygeno.2017.06.007>
- Choy K.T., Wong A.Y., Kaewpreedee P., Sia S.F., Chen D., Hui K., Chu D., Chan M., Cheung P.P., Huang X., Peiris M., Yen H.L. Remdesivir, lopinavir, emetine, and homoharringtonine inhibit SARS-CoV-2 replication in vitro. *Antivir. Res.* 2020;178 doi: 10.1016/j.antiviral.2020.104786.
- Clark D. E. (2006). What has computer-aided molecular design ever done for drug discovery?. *Expert opinion on drug discovery*, 1(2), 103–110. <https://doi.org/10.1517/17460441.1.2.103>
- Clark L.K., Green T.J., Petit C.M. Structure of nonstructural protein 1 from SARS-CoV-2. *bioRxiv. Preprint Server Biol.* 2020;2020 doi: 10.1101/2020.11.03.366757. 11.03.366757.

- Coopersmith C.M., Antonelli M., Bauer S.R., Deutschman C.S., Evans L.E., Ferrer R., Hellman J., Jog S., Kesecioglu J., Kissoon N., Martin-Loeches I., Nunnally M.E., Prescott H.C., Rhodes A., Talmor D., Tissieres P., De Backer D. The surviving sepsis campaign: research priorities for coronavirus disease 2019 in critical illness. *Crit. Care Med.* 2021;49(4):598–622. doi: 10.1097/CCM.0000000000004895.
- Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C (2009) Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J Chem Inf Model* 49:1455–1474
- Cui J., Li F., Shi Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* 2019;17:181–192.
- Dadashpour, S., Tuylu Kucukkilinc, T., Unsal Tan, O., Ozadali, K., Irannejad, H., & Emami, S. (2015). Design, synthesis and in vitro study of 5,6-diaryl-1,2,4-triazine-3-ylthioacetate derivatives as COX-2 and β -amyloid aggregation inhibitors. *Archiv der Pharmazie*, 348(3), 179–187. <https://doi.org/10.1002/ardp.201400400>
- Daugelaite, J., O' Driscoll, A., & Sleator, R. D. (2013). An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics. *International Scholarly Research Notices*, 2013(1), 615630. <https://doi.org/10.1155/2013/615630>
- Davies J.P., Almasy K.M., McDonald E.F., Plate L. Comparative multiplexed interactomics of SARS-CoV-2 and homologous coronavirus non-structural proteins identifies unique and shared host-cell dependencies. *bioRxiv* : Preprint Server Biol. 2020;2020doi: 10.1101/2020.07.13.201517. 07.13.201517.
- Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A Model of Evolutionary Change in Proteins. *Atlas of Protein Sequence and Structure*, 5, 345-351.
- Daze, K. and Hof, F. (2016). Molecular Interaction and Recognition. In *Encyclopedia of Physical Organic Chemistry* (eds Z. Wang, U. Wille and E. Juaristi). <https://doi.org/10.1002/9781118468586.epoc3001>
- De Meyer S., Bojkova D., Cinatl J., Van Damme E., Buyck C., Van Loock M., Woodfall B., Ciesek S. Lack of antiviral activity of darunavir against SARS-CoV-2. *Int. J. Infect. Dis.* 2020;97:7–10. doi: 10.1016/j.ijid.2020.05.085.
- Decroly E., Debarnot C., Ferron F., Bouvet M., Coutard B., Imbert I., Gluais L., Papageorgiou N., Sharff A., Bricogne G., Ortiz-Lombardia M., Lescar J., Canard B. Crystal structure and functional analysis of the SARS-coronavirus RNA cap 2'-O-methyltransferase nsp10/nsp16 complex. *PLoS Pathog.* 2011;7(5) doi: 10.1371/journal.ppat.1002059.
- Degorce, F. (2009). HTRF: A Technology Tailored for Drug Discovery - A Review of Theoretical Aspects and Recent Applications. *Current Chemical Genomics*, 3(1), 22–32. <https://doi.org/10.2174/1875397300903010022>
- Delang, L., Segura Guerrero, N., Tas, A., Quérat, G., Pastorino, B., Froeyen, M., Dallmeier, K., Jochmans, D., Herdewijn, P., Bello, F., Snijder, E. J., de Lamballerie, X., Martina, B., Neyts, J., van Hemert, M. J., & Leyssen, P. (2014). Mutations in the chikungunya virus non-structural proteins cause resistance to favipiravir (T-705), a broad-spectrum antiviral. *Journal of Antimicrobial Chemotherapy*, 69(10), 2770–2784. <https://doi.org/10.1093/jac/dku209>

- Dhankhar P., Dalal V., Kumar V. Screening of severe acute respiratory syndrome coronavirus 2 RNA-dependent RNA polymerase inhibitors using computational approach. *J. Comput. Biol.* 2021;28(12):1228–1247. doi: 10.1089/cmb.2020.0639.
- Du, X., Li, Y., Xia, Y. L., Ai, S. M., Liang, J., Sang, P., Ji, X. L., & Liu, S. Q. (2016). Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. *International journal of molecular sciences*, 17(2), 144. <https://doi.org/10.3390/ijms17020144>
- Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis*. Cambridge University Press; 1998.
- Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC biology*, 9, 71. <https://doi.org/10.1186/1741-7007-9-71>
- Eastman, P., Galvelis, R., Peláez, R. P., Abreu, C. R. A., Farr, S. E., Gallicchio, E., Gorenko, A., Henry, M. M., Hu, F., Huang, J., Krämer, A., Michel, J., Mitchell, J. A., Pande, V. S., Rodrigues, J. P., Rodriguez-Guerra, J., Simmonett, A. C., Singh, S., Swails, J., ... Markland, T. E. (2024). OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *The Journal of Physical Chemistry B*, 128(1), 109–116. <https://doi.org/10.1021/acs.jpcb.3c06662>
- Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1), 113. <https://doi.org/10.1186/1471-2105-5-113>
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), 1792–1797.
- Egloff M.P., Ferron F., Campanacci V., Longhi S., Rancurel C., Dutartre H., Snijder E.J., Gorbalenya A.E., Cambillau C., Canard B. The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. *Proc. Natl. Acad. Sci. U.S.A.* 2004;101(11):3792–3796. doi: 10.1073/pnas.0307877101.
- Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., & Mee, R. P. (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design*, 11(5), 425–445. <https://doi.org/10.1023/A:1007996124545>
- Elfiky A.A., Azzam E.B. Novel guanosine derivatives against MERS CoV polymerase: an in silicoperspective. *J. Biomol. Struct. Dynam.* 2021;39(8):2923–2931. doi: 10.1080/07391102.2020.1758789.
- Fayne D. De-Peptidising Protein-Protein Interactions – big jobs for small molecules. *Drug Discov. Today Technol.* 2013;10(4):e467–e474.
- Fehr A.R., Perlman S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.* 2015;1282:1–23. doi: 10.1007/978-1-4939-2438-7_1.
- Fehr, A. R., Channappanavar, R., Jankevicius, G., Fett, C., Zhao, J., Athmer, J., Meyerholz, D. K., Ahel, I., & Perlman, S. (2016). The Conserved Coronavirus Macrodome Promotes Virulence and Suppresses the Innate Immune Response during Severe Acute Respiratory Syndrome Coronavirus Infection. *mBio*, 7(6), e01721–16. <https://doi.org/10.1128/mBio.01721-16>
- Fischer, M., & Staples, J. E. (2014). Chikungunya Virus Spreads in the Americas—Caribbean and South America, 2013–2014. *Morbidity and Mortality Weekly Report*, 63(22), 500–501.

- Freitas B.T., Durie I.A., Murray J., Longo J.E., Miller H.C., Crich D., Hogan R.J., Tripp R.A., Pegan S.D. Characterization and noncovalent inhibition of the deubiquitinase and deISGylase activity of SARS-CoV-2 papain-like protease. *ACS Infect. Dis.* 2020;6(8):2099–2109. doi: 10.1021/acsinfecdis.0c00168.
- Frolov, I., & Frolova, E. I. (2022). Molecular Virology of Chikungunya Virus. In M. Heise (Ed.), *Chikungunya Virus* (pp. 1–31). Springer International Publishing. https://doi.org/10.1007/82_2018_146
- Gadhav, K., Kumar, P., Kumar, A., Bhardwaj, T., Garg, N., & Giri, R. (2021). Conformational dynamics of 13 amino acids long NSP11 of SARS-CoV-2 under membrane mimetics and different solvent conditions. *Microbial pathogenesis*, 158, 105041. <https://doi.org/10.1016/j.micpath.2021.105041>
- Gahbauer, S., Correy, G. J., Schuller, M., Ferla, M. P., Doruk, Y. U., Rachman, M., Wu, T., Diolaiti, M., Wang, S., Neitz, R. J., Fearon, D., Radchenko, D. S., Moroz, Y. S., Irwin, J. J., Renslo, A. R., Taylor, J. C., Gestwicki, J. E., von Delft, F., Ashworth, A., Ahel, I., ... Fraser, J. S. (2023). Iterative computational design and crystallographic screening identifies potent inhibitors targeting the Nsp3 macrodomain of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America*, 120(2), e2212931120. <https://doi.org/10.1073/pnas.2212931120>
- Gao Y., Yan L., Huang Y., Liu F., Zhao Y., Cao L., Wang T., Sun Q., Ming Z., Zhang L., Ge J., Zheng L., Zhang Y., Wang H., Zhu Y., Zhu C., Hu T., Hua T., Zhang B., Yang X., et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science (New York, N.Y.)* 2020;368(6492):779–782. doi: 10.1126/science.abb7498.
- Ge X.Y., Li J.L., Yang X.L., Chmura A.A., Zhu G., Epstein J.H., et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*. 2013;503:535–538.
- Gelderblom H.R. In: *Medical Microbiology*. fourth ed. Baron S., editor. University of Texas Medical Branch at Galveston; Galveston (TX): 1996. Structure and classification of viruses. <https://www.ncbi.nlm.nih.gov/books/NBK8174/> (Chapter 41). Available from:
- Gerlach P., Malet H., Cusack S., Reguera J. Structural insights into bunyavirus replication and its regulation by the vRNA promoter. *Cell*. 2015;161(6):1267–1279. doi: 10.1016/j.cell.2015.05.006.
- Ghahremanpour MM, Saar A, Tirado-Rives J, Jorgensen WL. Computation of Absolute Binding Free Energies for Noncovalent Inhibitors with SARS-CoV-2 Main Protease. *J Chem Inf Model*. 2023 Aug 28;63(16):5309–5318. doi: 10.1021/acs.jcim.3c00874. Epub 2023 Aug 10. PMID: 37561001.
- Ghosh A.K., Brindisi M., Shahabi D., Chapman M.E., Mesecar A.D. Drug development and medicinal chemistry efforts toward SARS-coronavirus and covid-19 therapeutics. *ChemMedChem*. 2020;15(11):907–932. doi: 10.1002/cmdc.202000223.
- Giordano D, Biancaniello C, Argenio MA, Facchiano A. Drug Design by Pharmacophore and Virtual Screening Approach. *Pharmaceuticals (Basel)*. 2022 May 23;15(5):646. doi: 10.3390/ph15050646. PMID: 35631472; PMCID: PMC9145410.
- Godoy, A. S., Nakamura, A. M., Douangamath, A., Song, Y., Noske, G. D., Gawriljuk, V. O., Fernandes, R. S., Pereira, H. D. M., Oliveira, K. I. Z., Fearon, D., Dias, A., Krojer, T., Fairhead, M., Powell, A., Dunnet, L., Brandao-Neto, J., Skyner, R., Chalk, R., Bajusz, D., Bege, M., ... Oliva, G. (2023). Allosteric regulation and crystallographic fragment screening of SARS-CoV-2 NSP15

- endoribonuclease. *Nucleic acids research*, 51(10), 5255–5270. <https://doi.org/10.1093/nar/gkad314>
- Gonnet GH, Cohen MA, Benner SA: Exhaustive matching of the entire protein sequence database. *Science* 1992, 256(5062):1443–1445.
- Gorbalenya A.E. Big nidovirus genome. When count and order of domains matter. *Adv. Exp. Med. Biol.* 2001;494:1–17.
- Gorbalenya A.E., Enjuanes L., Ziebuhr J., Snijder E.J. Nidovirales: evolving the largest RNA virus genome. *Virus Res.* 2006;117(1):17–37. doi: 10.1016/j.virusres.2006.01.017.
- Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M. Coronavirus genome: prediction of putative functional domains in the non- structural polyprotein by comparative amino acid sequence analysis. *Nucleic Acids Res.* 1989;17:4847–4861. doi: 10.1093/nar/17.12.4847.
- Gorbalenya A.E., Koonin E.V., Lai M.M. Putative papain- related thiol proteases of positive-strand RNA viruses. Identification of rubi- and aphthovirus proteases and delineation of a novel conserved domain associated with proteases of rubi-, alpha- and coronaviruses. *FEBS Lett.* 1991;288:201–205.
- Gordon D.E., Jang G.M., Bouhaddou M., Xu J., Obernier K., White K.M., O'Meara M.J., Rezelj V.V., Guo J.Z., Swaney D.L., Tummino T.A., Hüttenhain R., Kaake R.M., Richards A.L., Tutuncuoglu B., Foussard H., Batra J., Haas K., Modak M., Kim M., et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*. 2020;583(7816):459–468. doi: 10.1038/s41586-020-2286-9.
- Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3), 705–708. [https://doi.org/10.1016/0022-2836\(82\)90398-9](https://doi.org/10.1016/0022-2836(82)90398-9)
- Gotoh, O. (1990). Consistency of optimal sequence alignments. *Bulletin of Mathematical Biology*, 52(4), 509–525. [https://doi.org/10.1016/S0092-8240\(05\)80359-3](https://doi.org/10.1016/S0092-8240(05)80359-3)
- Gotoh, O. (1996). Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments. *Journal of Molecular Biology*, 264(4), 823–838. <https://doi.org/10.1006/jmbi.1996.0679>
- Goyal B., Goyal D. Targeting the dimerization of the main protease of coronaviruses: a potential broad-spectrum therapeutic strategy. *ACS Comb. Sci.* 2020;22(6):297–305. doi: 10.1021/acscombsci.0c00058.
- Graham R.L., Donaldson E.F., Baric R.S. A decade after SARS: strategies for controlling emerging coronaviruses. *Nat. Rev. Microbiol.* 2013;11:836–848.
- Grant, J. A., & Pickup, B. T. (1995). A Gaussian Description of Molecular Shape. *The Journal of Physical Chemistry*, 99(11), 3503–3510. <https://doi.org/10.1021/j100011a016>
- Grant, J. A., Gallardo, M. A., & Pickup, B. T. (1996). A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry*, 17(14), 1653–1666. [https://doi.org/10.1002/\(SICI\)1096-987X\(19961115\)17:14<1653::AID-JCC7>3.0.CO;2-K](https://doi.org/10.1002/(SICI)1096-987X(19961115)17:14<1653::AID-JCC7>3.0.CO;2-K)
- Gronau, I., & Moran, S. (2007). Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters*, 104(6), 205–210.

- Grover, S., Apushkin, M. A., & Fishman, G. A. (2006). Topical dorzolamide for the treatment of cystoid macular edema in patients with retinitis pigmentosa. *American journal of ophthalmology*, 141(5), 850–858. <https://doi.org/10.1016/j.ajo.2005.12.030>
- Guan W., Ni Z., Hu Y., et al. Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* 2020;58(4):711–712. doi: 10.1016/j.jemermed.2020.04.004. [Online ahead of print] DOI: 10.1056/NEJMoa2002032. The Journal of Emergency Medicine.
- Guerrero-Arguero, I., Tellez-Freitas, C. M., Weber, K. S., Berges, B. K., Robison, R. A., & Pickett, B. E. (2021). Alphaviruses: Host pathogenesis, immune response, and vaccine & treatment updates. *Journal of General Virology*, 102(8), 001644. <https://doi.org/10.1099/jgv.0.001644>
- Gupta R., Charron J., Stenger C.L., Painter J., Steward H., Cook T.W., Faber W., Frisch A., Lind E., Bauss J., Li X., Sirpilla O., Soehnlén X., Underwood A., Hinds D., Morris M., Lamb N., Carcillo J.A., Bupp C., Uhal B.D., et al. SARS-CoV-2 (COVID-19) structural and evolutionary dynamicome: insights into functional evolution and human genomics. *J. Biol. Chem.* 2020;295(33):11742–11753. doi: 10.1074/jbc.RA120.014873.
- Gupta, D. K., Kaur, P., Leong, S. T., Tan, L. T., Prinsep, M. R., & Chu, J. J. H. (2014). Anti-Chikungunya Viral Activities of Aplysiatoxin-Related Compounds from the Marine Cyanobacterium *Trichodesmium erythraeum*. *Marine Drugs*, 12(1), Article 1. <https://doi.org/10.3390/md12010115>
- Haider, K. (2010). Computational analyses of protein-ligand interactions.
- Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., & Hutchison, G. R. (2012). Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of cheminformatics*, 4(1), 17. <https://doi.org/10.1186/1758-2946-4-17>
- Harcourt B.H., Jukneliene D., Kanjanahaluethai A., Bechill J., Severson K.M., Smith C.M., Rota P.A., Baker S.C. Identification of severe acute respiratory syndrome coronavirus replicase products and characterization of papain-like protease activity. *J. Virol.* 2004;78(24):13600–13612. doi: 10.1128/JVI.78.24.13600-13612.2004.
- Hawkins PC, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. *J Med Chem.* 2007 Jan 11;50(1):74-82. doi: 10.1021/jm0603365. PMID: 17201411.
- He X, Walker B, Man VH, Ren P, Wang J. Recent progress in general force fields of small molecules. *Curr Opin Struct Biol.* 2022 Feb;72:187-193. doi: 10.1016/j.sbi.2021.11.011. Epub 2021 Dec 20. PMID: 34942567; PMCID: PMC8860847.
- Heilmann E., Costacurta F., Moghadas S.A., Ye C., Pavan M., Bassani D., Volland A., Ascher C., Weiss A.K.H., Bante D., Harris R.S., Moro S., Rupp B., Martinez-Sobrido L., von Laer D. SARS-CoV-2 3CL^{pro} mutations selected in a VSV-based system confer resistance to nirmatrelvir, ensitrelvir, and GC376. *Sci. Transl. Med.* 2023;15(678) doi: 10.1126/scitranslmed.abq7360.
- Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22), 10915–10919. <https://doi.org/10.1073/pnas.89.22.10915>
- Herasymenko O, Silva M, Abu-Saleh AAA, Ahmad A, Alvarado-Huayhuaz J, Arce OEA, Armstrong RJ, Arrowsmith C, Bachta KE, Beck H, Berta D, Bieniek MK, Blay V, Bolotokova A, Bourne PE, Breznik M, Brown PJ, Campbell ADG, Carosati E, Chau I, Cole DJ, Cree B, Dehaen W, Denzinger K, Dos Santos Machado K, Dunn I, Durai P, Edfeldt K, Edwards A, Fayne D, Felfoldi D, Friston K,

- Ghiabi P, Gibson E, Günther J, Gunnarsson A, Hillisch A, Houston DR, Jensen JH, Harding RJ, Harris KS, Hoffer L, Hogner A, Horton JT, Houlston S, Hultquist JF, Hutchinson A, Irwin JJ, Jukić M, Kandwal S, Karlova A, Katis VL, Kich RP, Kireev D, Koes D, Inniss NL, Lessel U, Liu S, Loppnau P, Lu W, Martino S, McGibbon M, Meiler J, Mettu A, Money-Kyrle S, Moretti R, Moroz YS, Muvva C, Newman JA, Obendorf L, Paige B, Pandit A, Park K, Perveen S, Pirie R, Poda G, Protopopov M, Pütter V, Ricci F, Roper NJ, Rosta E, Rzhetskaya M, Sabnis Y, Satchell KJF, Schmitt Kremer F, Scott T, Seitova A, Steinmann C, Talagayev V, Tarkhanova OO, Tatum NJ, Treleaven D, Velasque Werhli A, Walters WP, Wang X, Wells J, Wells G, Westermaier Y, Wolber G, Wortmann L, Zhang J, Zhao Z, Zheng S, Schapira M. CACHE Challenge #2: Targeting the RNA Site of the SARS-CoV-2 Helicase Nsp13. *J Chem Inf Model*. 2025 Jul 14;65(13):6884-6898. doi: 10.1021/acs.jcim.5c00535. Epub 2025 Jun 20. PMID: 40539604; PMCID: PMC12323591.
- Herasymenko, O., Silva, M., Correy, G. J., Abu-Saleh, A. A.-A. A., Ackloo, S., Arrowsmith, C., Ban, F., Beck, H., Bishop, K. P., Bohórquez, H. J., Bolotokova, A., Breznik, M., Chau, I., Chen, Y., Cherkasov, A., Dehaen, W., Corte, D. D., Denzinger, K., Edfeldt, K., ... Schapira, M. (2025). *CACHE Challenge #3: Targeting the Nsp3 Macrodomein of SARS-CoV-2*. ChemRxiv. <https://doi.org/10.26434/chemrxiv-2025-d1v7v>
- Hess, B., Bekker, H., Berendsen, H. J. C., & Fraaije, J. G. E. M. (1997). LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12), 1463–1472. [https://doi.org/10.1002/\(SICI\)1096-987X\(199709\)18:12<1463::AID-JCC4>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H)
- Higashi-Kuwata N, Bulut H, Hayashi H, Tsuji K, Ogata-Aoki H, Kiso M, Takamune N, Kishimoto N, Hattori SI, Ishii T, Kobayakawa T, Nakano K, Shimizu Y, Das D, Saruwatari J, Hasegawa K, Murayama K, Sukenaga Y, Takamatsu Y, Yoshimura K, Aoki M, Furusawa Y, Okamura T, Yamayoshi S, Kawaoka Y, Misumi S, Tamamura H, Mitsuya H. An orally available P1'-5-fluorinated M^{pro} inhibitor blocks SARS-CoV-2 replication without booster and exhibits high genetic barrier. *PNAS Nexus*. 2025 Jan 7;4(1):pgae578. doi: 10.1093/pnasnexus/pgae578. PMID: 39831159; PMCID: PMC11740726.
- Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, 99(6), 1129–1143. <https://doi.org/10.1016/j.neuron.2018.08.011>
- Holm L, Sander C. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res*. 1999;27(1):244–247. doi: 10.1093/nar/27.1.244.
- Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol*. 1993;233(1):123–138. doi: 10.1006/jmbi.1993.1489.
- Holmes, A. C., Basore, K., Fremont, D. H., & Diamond, M. S. (2020). A molecular understanding of alphavirus entry. *PLOS Pathogens*, 16(10), e1008876. <https://doi.org/10.1371/journal.ppat.1008876>
- Hospital, A., Goñi, J. R., Orozco, M., & Gelpí, J. L. (2015). Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, 8, 37–47. <https://doi.org/10.2147/AABC.S70333>
- Huang C., Wang Y., Li X., Ren L., Zhao J., Hu Y., Zhang L., Fan G., Xu J., Gu X., Cheng Z., Yu T., Xia J., Wei Y., Wu W., Xie X., Yin W., Li H., Liu M., Xiao Y., et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London, England)* 2020;395(10223):497–506. doi: 10.1016/S0140-6736(20)30183-5.
- Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, MacKerell AD Jr. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat*

- Methods. 2017 Jan;14(1):71-73. doi: 10.1038/nmeth.4067. Epub 2016 Nov 7. PMID: 27819658; PMCID: PMC5199616.
- Huang X., Dong W., Milewska A., Golda A., Qi Y., Zhu Q.K., Marasco W.A., Baric R.S., Sims A.C., Pyrc K., Li W., Sui J. Human coronavirus HKU1 spike protein uses O-acetylated sialic acid as an attachment receptor determinant and employs hemagglutinin-esterase protein as a receptor-destroying enzyme. *J. Virol.* 2015;89(14):7202–7213. doi: 10.1128/JVI.00854-15.
- Hucke, F. I. L., & Bugert, J. J. (2020). Current and Promising Antivirals Against Chikungunya Virus. *Frontiers in Public Health*, 8, 618624. <https://doi.org/10.3389/fpubh.2020.618624>
- Hui D.S., I Azhar E., Madani T.A., Ntoumi F., Kock R., Dar O., Ippolito G., Mchugh T.D., Memish Z.A., Drosten C., Zumla A., Petersen E. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - the latest 2019 novel coronavirus outbreak in Wuhan, China. *Int. J. Infect. Dis.* 2020;91:264–266. doi: 10.1016/j.ijid.2020.01.009.
- Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38.
- Ibrahim, M. K., Yusof, U. K., Eisa, T. A. E., & Nasser, M. (2024). Bioinspired Algorithms for Multiple Sequence Alignment: A Systematic Review and Roadmap. *Applied Sciences*, 14(6), Article 6. <https://doi.org/10.3390/app14062433>
- ICTV (International Committee on Taxonomy of Virus) ICTV; 2019. Virus Taxonomy: 2019 Release.
- Jana I.D., Bhattacharya P., Mayilsamy K., Banerjee S., Bhattacharje G., Das S., Aditya S., Ghosh A., McGill A.R., Srikrishnan S., Das A.K., Basak A., Mohapatra S.S., Chandran B., Bhimsaria D., Mohapatra S., Roy A., Mondal A. Targeting an evolutionarily conserved "E-L-L" motif in the spike protein to develop a small molecule fusion inhibitor against SARS-CoV-2. *bioRxiv.* 2022;16doi: 10.1101/2022.03.16.484554.PMID:35313575. [Preprint]. 2022.03. PMCID: PMC8936095.
- Jang W.D., Jeon S., Kim S., Lee S.Y. Drugs repurposed for COVID-19 by virtual screening of 6,218 drugs and cell-based assay. *Proc. Natl. Acad. Sci. U.S.A.* 2021;118(30) doi: 10.1073/pnas.2024302118.
- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2), 926–935. <https://doi.org/10.1063/1.445869>
- Jose, J., Snyder, J. E., & Kuhn, R. J. (2009). A structural and functional perspective of alphavirus replication and assembly. *Future Microbiology*, 4, 837–856. <https://doi.org/10.2217/fmb.09.59>
- Joseph J.S., Saikatendu K.S., Subramanian V., Neuman B.W., Brooun A., Griffith M., Moy K., Yadav M.K., Velasquez J., Buchmeier M.J., Stevens R.C., Kuhn P. Crystal structure of nonstructural protein 10 from the severe acute respiratory syndrome coronavirus reveals a novel fold with two zinc-binding motifs. *J. Virol.* 2006;80(16):7894–7901. doi: 10.1128/JVI.00467-06.
- Joseph J.S., Saikatendu K.S., Subramanian V., Neuman B.W., Buchmeier M.J., Stevens R.C., Kuhn P. Crystal structure of a monomeric form of severe acute respiratory syndrome coronavirus endonuclease nsp15 suggests a role for hexamerization as an allosteric switch. *J. Virol.* 2007;81(12):6700–6708. doi: 10.1128/JVI.02817-06.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021 Aug;596(7873):583–589. doi: 10.1038/s41586-021-03819-2. Epub 2021 Jul 15. PMID: 34265844; PMCID: PMC8371605.

Kalnins, G., Rudusa, L., Bula, A. L., Zelencova-Gopejenko, D., Bobileva, O., Sisovs, M., Tars, K., Jirgensons, A., Jaudzems, K., & Bobrovs, R. (2024). Structural Basis for Inhibition of the SARS-CoV-2 nsp16 by Substrate-Based Dual Site Inhibitors. *ChemMedChem*, 19(24), e202400618. <https://doi.org/10.1002/cmdc.202400618>

Kalyanamoorthy, S., & Chen, Y. P. (2011). Structure-based drug design to augment hit discovery. *Drug discovery today*, 16(17-18), 831–839. <https://doi.org/10.1016/j.drudis.2011.07.006>

Kandwal S., Fayne D. Repurposing drugs for treatment of SARS-CoV-2 infection: computational design insights into mechanisms of action. *J. Biomol. Struct. Dynam.* 2022;40(3):1316–1330. doi: 10.1080/07391102.2020.1825232.

Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9), 646–652. <https://doi.org/10.1038/nsb0902-646>

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, 9(4), 286–298. <https://doi.org/10.1093/bib/bbn013>

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059–3066. <https://doi.org/10.1093/nar/gkf436>

Kendall, C., Khalid, H., Müller, M., Banda, D. H., Kohl, A., Merits, A., Stonehouse, N. J., & Tuplin, A. (2019). Structural and phenotypic analysis of Chikungunya virus RNA replication elements. *Nucleic Acids Research*, 47(17), 9296–9312. <https://doi.org/10.1093/nar/gkz640>

Khalak, Y., Tresadern, G., Aldeghi, M., Baumann, H. M., Mobley, D. L., Groot, B. L. de, & Gapsys, V. (2021). Alchemical absolute protein–ligand binding free energies for drug design. *Chemical Science*, 12(41), 13958–13971. <https://doi.org/10.1039/D1SC03472C>

Khan MF, Kandwal S, Fayne D. DataPype: A Fully Automated Unified Software Platform for Computer-Aided Drug Design. *ACS Omega*. 2023 Oct 12;8(42):39468–39480. doi: 10.1021/acsomega.3c05207. PMID: 37901539; PMCID: PMC10601415.

Khan, M., Santhosh, S. R., Tiwari, M., Lakshmana Rao, P. V., & Parida, M. (2010). Assessment of in vitro prophylactic and therapeutic efficacy of chloroquine against chikungunya virus in vero cells. *Journal of Medical Virology*, 82(5), 817–824. <https://doi.org/10.1002/jmv.21663>

Khare S., Gurry C., Freitas L., Schultz M.B., Bach G., Diallo A., Akite N., Ho J., Lee R.T., Yeo W., Curation Team G.C., Maurer-Stroh S. GISAID's role in pandemic response. *Chin CDC Week*. 2021;3(49):1049–1051. doi: 10.46234/ccdcw2021.255.

Kim D., Lee J.Y., Yang J.S., Kim J.W., Kim V.N., Chang H. The architecture of SARS-CoV-2 transcriptome. *Cell*. 2020;181(4):914–921.e10. doi: 10.1016/j.cell.2020.04.011.

Kim M.K., Yu M.S., Park H.R., Kim K.B., Lee C., Cho S.Y., Kang J., Yoon H., Kim D.E., Choo H., Jeong Y.J., Chong Y. 2,6-Bis-aryl-methoxy-5-hydroxychromones with antiviral activity against both

- hepatitis C virus (HCV) and SARS-associated coronavirus (SCV) *Eur. J. Med. Chem.* 2011;46(11):5698–5704. doi: 10.1016/j.ejmech.2011.09.005.
- Kim Y., Jedrzejczak R., Maltseva N.I., Wilamowski M., Endres M., Godzik A., Michalska K., Joachimiak A. Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Protein Sci.* 2020;29(7):1596–1605. doi: 10.1002/pro.3873.
- Kim, Y., Wower, J., Maltseva, N., Chang, C., Jedrzejczak, R., Wilamowski, M., Kang, S., Nicolaescu, V., Randall, G., Michalska, K., & Joachimiak, A. (2021). Tipiracil binds to uridine site and inhibits Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Communications biology*, 4(1), 193. <https://doi.org/10.1038/s42003-021-01735-9>
- Kirchdoerfer R.N., Ward A.B. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat. Commun.* 2019;10(1):2342. doi: 10.1038/s41467-019-10280-3.
- Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., & Cheatham, T. E., 3rd (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research*, 33(12), 889–897. <https://doi.org/10.1021/ar000033j>
- Koonin E.V., Gorbalenya A.E., Chumakov K.M. Tentative identification of RNA-dependent RNA polymerases of dsRNA viruses and their relationship to positive strand RNA viral polymerases. *FEBS (Fed. Eur. Biochem. Soc.) Lett.* 1989;252(1–2):42–46. doi: 10.1016/0014-5793(89)80886-5.
- Krafcikova P., Silhan J., Nencka R., Boura E. Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. *Nat. Commun.* 2020;11(1):3717. doi: 10.1038/s41467-020-17495-9.
- Kubes, V., & Ríos, F. A. (1939). The Causative Agent of Infectious Equine Encephalomyelitis in Venezuela. *Science*, 90(2323), 20–21. <https://doi.org/10.1126/science.90.2323.20>
- Kumar, A., & MacKerell, A. D., Jr (2024). FFParm-v2.0: A Comprehensive Tool for CHARMM Additive and Drude Polarizable Force-Field Parameter Optimization and Validation. *The journal of physical chemistry. B*, 128(18), 4385–4395. <https://doi.org/10.1021/acs.jpcc.4c01314>
- Kumar, A., & Zhang, K. Y. J. (2018). Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Frontiers in chemistry*, 6, 315. <https://doi.org/10.3389/fchem.2018.00315>
- Kumar, S., & Filipski, A. (2007). Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome research*, 17(2), 127–135.
- Kuri, T., Eriksson, K. K., Putics, A., Züst, R., Snijder, E. J., Davidson, A. D., Siddell, S. G., Thiel, V., Ziebuhr, J., & Weber, F. (2011). The ADP-ribose-1"-monophosphatase domains of severe acute respiratory syndrome coronavirus and human coronavirus 229E mediate resistance to antiviral interferon responses. *The Journal of general virology*, 92(Pt 8), 1899–1905. <https://doi.org/10.1099/vir.0.031856-0>
- Kutzner C, Páll S, Fechner M, Esztermann A, de Groot BL, Grubmüller H. Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *J Comput Chem.* 2015 Oct 5;36(26):1990–2008. doi: 10.1002/jcc.24030. Epub 2015 Aug 4. PMID: 26238484; PMCID: PMC5042102.

- Kwong A.D., Rao B.G., Jeang K.T. Viral and cellular RNA helicases as antiviral targets. *Nat. Rev. Drug Discov.* 2005;4(10):845–853. doi: 10.1038/nrd1853.
- LaPlante, S. R., Padyana, A. K., Abeywardane, A., Bonneau, P., Cartier, M., Coulombe, R., Jakalian, A., Wildeson-Jones, J., Li, X., Liang, S., McKercher, G., White, P., Zhang, Q., & Taylor, S. J. (2014). Integrated strategies for identifying leads that target the NS3 helicase of the hepatitis C virus. *Journal of medicinal chemistry*, 57(5), 2074–2090. <https://doi.org/10.1021/jm401432c>
- Lassmann, T., & Sonnhammer, E. L. (2005). Kalign – an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6(1), 298. <https://doi.org/10.1186/1471-2105-6-298>
- Lau E.Y., Negrete O.A., Bennett W., Bennion B.J., Borucki M., Bourguet F., Epstein A., Franco M., Harmon B., He S., Jones D., Kim H., Kirshner D., Lao V., Lo J., McLoughlin K., Mosesso R., Murugesh D.K., Saada E.A., Segelke B., et al. Discovery of small-molecule inhibitors of SARS-CoV-2 proteins using a computational and experimental pipeline. *Front. Mol. Biosci.* 2021;8doi: 10.3389/fmolb.2021.678701.
- Lau Y.L., Peiris J.M. Pathogenesis of severe acute respiratory syndrome. *Curr. Opin. Immunol.* 2005;17:404–410.
- Laurent, T., Kumar, P., Liese, S., Zare, F., Jonasson, M., Carlson, A., & Carlson, L.-A. (2022). Architecture of the chikungunya virus replication organelle. *eLife*, 11, e83042. <https://doi.org/10.7554/eLife.83042>
- Lavigne M., Helynck O., Rigolet P., Boudria-Souilah R., Nowakowski M., Baron B., Brûlé S., Hoos S., Raynal B., Guittat L., Beauvineau C., Petres S., Granzhan A., Guillon J., Pratviel G., Teulade-Fichou M.P., England P., Mergny J.L., Munier-Lehmann H. SARS-CoV-2 Nsp3 unique domain SUD interacts with guanine quadruplexes and G4-ligands inhibit this interaction. *Nucleic Acids Res.* 2021;49(13):7695–7712. doi: 10.1093/nar/gkab571.
- Lee C., Lee J.M., Lee N.R., Kim D.E., Jeong Y.J., Chong Y. Investigation of the pharmacophore space of Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) NTPase/helicase by dihydroxychromone derivatives. *Bioorg. Med. Chem. Lett.* 2009;19(16):4538–4541. doi: 10.1016/j.bmcl.2009.07.009.
- Littler D.R., Gully B.S., Colson R.N., Rossjohn J. Crystal structure of the SARS-CoV-2 non-structural protein 9, Nsp9. *iScience*. 2020;23(7) doi: 10.1016/j.isci.2020.101258.
- Liu, K., & Kokubo, H. (2020). Prediction of ligand binding mode among multiple cross-docking poses by molecular dynamics simulations. *Journal of computer-aided molecular design*, 34(11), 1195–1205. <https://doi.org/10.1007/s10822-020-00340-y>
- Liu X., Li Z., Liu S., Sun J., Chen Z., Jiang M., Zhang Q., Wei Y., Wang X., Huang Y.Y., Shi Y., Xu Y., Xian H., Bai F., Ou C., Xiong B., Lew A.M., Cui J., Fang R., Huang H., et al. Potential therapeutic effects of dipyridamole in the severely ill patients with COVID-19. *Acta Pharm. Sin. B.* 2020;10(7):1205–1215. doi: 10.1016/j.apsb.2020.04.008.
- Liu, Y., Schmidt, B., & Maskell, D. L. (2010). MSAProbs: Multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics*, 26(16), 1958–1964. <https://doi.org/10.1093/bioinformatics/btq338>
- Ma Y., Wu L., Shaw N., Gao Y., Wang J., Sun Y., Lou Z., Yan L., Zhang R., Rao Z. Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proc. Natl. Acad. Sci. U.S.A.* 2015;112(30):9436–9441. doi: 10.1073/pnas.1508686112.

- MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., ... Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry. B*, 102(18), 3586–3616. <https://doi.org/10.1021/jp973084f>
- Madeira F., Pearce M., Tivey A., Basutkar P., Lee J., Edbali O., Madhusoodanan N., Kolesnikov A., Lopez R. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* 2022;50(W1):W276–W279. doi: 10.1093/nar/gkac240. Advance online publication.
- Manolaridis I., Wojdyla J.A., Panjikar S., Snijder E.J., Gorbalenya A.E., Berglund H., Nordlund P., Coutard B., Tucker P.A. Structure of the C-terminal domain of nsp4 from feline coronavirus. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 2009;65(Pt 8):839–846. doi: 10.1107/S0907444909018253.
- Marrakchi, H., Lanéelle, G., & Quémard, A. K. (2000). InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II. *Microbiology (Reading, England)*, 146 (Pt 2), 289–296. <https://doi.org/10.1099/00221287-146-2-289>
- McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature*, 267(5612), 585–590. <https://doi.org/10.1038/267585a0>
- McGann M. (2011). FRED pose prediction and virtual screening accuracy. *Journal of chemical information and modeling*, 51(3), 578–596. <https://doi.org/10.1021/ci100436p>
- MEGA11: Molecular Evolutionary Genetics Analysis version 11 (Tamura, Stecher, and Kumar 2021).
- Menachery V.D., Yount B.L., Jr., Debbink K., Agnihothram S., Gralinski L.E., Plante J.A., et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* 2015;21:1508–1513.
- Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2), 146–157. <https://doi.org/10.2174/157340911795677602>
- Mey, A. S. J. S., Allen, B. K., Macdonald, H. E. B., Chodera, J. D., Hahn, D. F., Kuhn, M., Michel, J., Mobley, D. L., Naden, L. N., Prasad, S., Rizzi, A., Scheen, J., Shirts, M. R., Tresadern, G., & Xu, H. (2020). Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *Living Journal of Computational Molecular Science*, 2(1), 18378. <https://doi.org/10.33011/livecoms.2.1.18378>
- Meyer, C., Garzia, A., Miller, M. W., Huggins, D. J., Myers, R. W., Hoffmann, H. H., Ashbrook, A. W., Jannath, S. Y., Liverton, N., Kargman, S., Zimmerman, M., Nelson, A. M., Sharma, V., Dolgov, E., Cangialosi, J., Penalva-Lopez, S., Alvarez, N., Chang, C. W., Oswal, N., Gonzalez, I., ... Tuschl, T. (2025). Small-molecule inhibition of SARS-CoV-2 NSP14 RNA cap methyltransferase. *Nature*, 637(8048), 1178–1185. <https://doi.org/10.1038/s41586-024-08320-0>
- Michalska K., Kim Y., Jedrzejczak R., Maltseva N.I., Stols L., Endres M., Joachimiak A. Crystal structures of SARS-CoV-2 ADP-ribose phosphatase: from the apo form to ligand complexes. *IUCrJ.* 2020;7(Pt 5):814–824. doi: 10.1107/S2052252520009653.
- Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem.* 2011 Jul 30;32(10):2319–27. doi: 10.1002/jcc.21787. Epub 2011 Apr 15. PMID: 21500218; PMCID: PMC3144279.

- Miknis Z.J., Donaldson E.F., Umland T.C., Rimmer R.A., Baric R.S., Schultz L.W. Severe acute respiratory syndrome coronavirus nsp9 dimerization is essential for efficient viral growth. *J. Virol.* 2009;83(7):3007–3018. doi: 10.1128/JVI.01505-08.
- Min Y.Q., Mo Q., Wang J., Deng F., Wang H., Ning Y.J. SARS-CoV-2 nsp1: bioinformatics, potential structural and functional features, and implications for drug/vaccine designs. *Front. Microbiol.* 2020;11 doi: 10.3389/fmicb.2020.587317.
- Modrow S., Falke D., Truyen U., Schätzl H. Viruses with single-stranded, positive-sense RNA genomes. *Mol. Virol.* 2013;185–349. doi: 10.1007/978-3-642-20718-1_14.
- Mohan, V., Gibbs, A. C., Cummings, M. D., Jaeger, E. P., & DesJarlais, R. L. (2005). Docking: successes and challenges. *Current pharmaceutical design*, 11(3), 323–333. <https://doi.org/10.2174/1381612053382106>
- Molecular Operating Environment (MOE) 2R7. 2022. p. 2022. (02 Chemical Computing Group ULC, 1010 Sherbooke St. West). Suite #910, Montreal, QC, Canada, H3A.
- Mönttinen H., Ravantti J.J., Poranen M.M. Structure unveils relationships between RNA virus polymerases. *Viruses*. 2021;13(2):313. doi: 10.3390/v13020313.
- Mount, D. W. Bioinformatics: sequence and genome analysis. 2004. *Bioinformatics: Sequence and Genome Analysis*.
- Murata, M., Richardson, J. S., & Sussman, J. L. (1985). Simultaneous comparison of three protein sequences. *Proceedings of the National Academy of Sciences*, 82(10), 3073–3077. <https://doi.org/10.1073/pnas.82.10.3073>
- Myers G. (1999) A fast bit-vector algorithm for approximate string matching based on dynamic programming . *J. ACM* , 46 , 395 –415
- Najjar-Debbiny R., Gronich N., Weber G., Khoury J., Amar M., Stein N., Goldstein L.H., Saliba W. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America; 2022. Effectiveness of Paxlovid in Reducing Severe COVID-19 and Mortality in High Risk Patients. ciac443. Advance online publication.
- Naqvi A., Fatima K., Mohammad T., Fatima U., Singh I.K., Singh A., Atif S.M., Hariprasad G., Hasan G.M., Hassan M.I. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: structural genomics approach. *Biochim. Biophys. Acta (BBA) - Mol. Basis Dis.* 2020;1866(10) doi: 10.1016/j.bbadis.2020.165878.
- Narayanan K., Huang C., Lokugamage K., Kamitani W., Ikegami T., Tseng C.T., Makino S. Severe acute respiratory syndrome coronavirus nsp1 suppresses host gene expression, including that of type I interferon, in infected cells. *J. Virol.* 2008;82(9):4471–4479. doi: 10.1128/JVI.02472-07.
- Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453. [https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4)
- Neuman B.W. Bioinformatics and functional analyses of coronavirus nonstructural proteins involved in the formation of replicative organelles. *Antivir. Res.* 2016;135:97–107. doi: 10.1016/j.antiviral.2016.10.005.

- Neuvonen M., Ahola T. Differential activities of cellular and viral macro domain proteins in binding of ADP-ribose metabolites. *J. Mol. Biol.* 2009;385(1):212–225. doi: 10.1016/j.jmb.2008.10.045.
- Newman, J. A., Douangamath, A., Yadzani, S., Yosaatmadja, Y., Aimon, A., Brandão-Neto, J., Dunnett, L., Gorrie-Stone, T., Skyner, R., Fearon, D., Schapira, M., von Delft, F., & Gileadi, O. (2021). Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nature communications*, 12(1), 4848. <https://doi.org/10.1038/s41467-021-25166-6>
- Nga P.T., Parquet Mdel C., Lauber C., Parida M., Nabeshima T., Yu F., et al. Discovery of the first insect nidovirus, a missing evolutionary link in the emergence of the largest RNA virus genomes. *PLoS Pathog.* 2011;7
- Nishikiori M., Sugiyama S., Xiang H., Niiyama M., Ishibashi K., Inoue T., Ishikawa M., Matsumura H., Katoh E. Crystal structure of the superfamily 1 helicase from Tomato mosaic virus. *J. Virol.* 2012;86(14):7565–7576. doi: 10.1128/JVI.00118-12.
- Notredame, C. (2002). Recent progress in multiple sequence alignment: A survey. *Pharmacogenomics*, 3(1), 131–144. <https://doi.org/10.1517/14622416.3.1.131>
- Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment1. *Journal of Molecular Biology*, 302(1), 205–217. <https://doi.org/10.1006/jmbi.2000.4042>
- O'Donnell, H. R., Tummino, T. A., Bardine, C., Craik, C. S., & Shoichet, B. K. (2021). Colloidal Aggregators in Biochemical SARS-CoV-2 Repurposing Screens. *Journal of medicinal chemistry*, 64(23), 17530–17539. <https://doi.org/10.1021/acs.jmedchem.1c01547>
- OpenEye Documentation <https://docs.eyesopen.com>
- Ou, J., Lan, W., Wu, X., Zhao, T., Duan, B., Yang, P., Ren, Y., Quan, L., Zhao, W., Seto, D., Chodosh, J., Luo, Z., Wu, J., & Zhang, Q. (2022). Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. *Signal Transduction and Targeted Therapy*, 7, 138. <https://doi.org/10.1038/s41392-022-00992-2>
- Ozden, S., Lucas-Hourani, M., Ceccaldi, P.-E., Basak, A., Valentine, M., Benjannet, S., Hamelin, J., Jacob, Y., Mamchaoui, K., Mouly, V., Desprès, P., Gessain, A., Butler-Browne, G., Chrétien, M., Tangy, F., Vidalain, P.-O., & Seidah, N. G. (2008). Inhibition of Chikungunya Virus Infection in Cultured Human Muscle Cells by Furin Inhibitors: IMPAIRMENT OF THE MATURATION OF THE E2 SURFACE GLYCOPROTEIN*. *Journal of Biological Chemistry*, 283(32), 21899–21908. <https://doi.org/10.1074/jbc.M802444200>
- Padhan K., Parvez M.K., Al-Dosari M.S. Comparative sequence analysis of SARS-CoV-2 suggests its high transmissibility and pathogenicity. *Future Virol.* 2021 doi: 10.2217/fvl-2020-0204.10.2217/fvl-2020-0204.
- Pant S., Singh M., Ravichandiran V., Murty U., Srivastava H.K. Peptide-like and small-molecule inhibitors against Covid-19. *J. Biomol. Struct. Dyn.* 2020;39(8):2904–2913. <https://www.tandfonline.com/doi/full/10.1080/07391102.2020.1757510>
- Parks J.M., Smith J.C. How to discover antiviral drugs quickly. *N. Engl. J. Med.* 2020;382(23):2261–2264. doi: 10.1056/NEJMcibr2007042.

- Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12), 7182–7190. <https://doi.org/10.1063/1.328693>
- Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T., & Barzilay, R. (2025). *Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction* (p. 2025.06.14.659707). bioRxiv. <https://doi.org/10.1101/2025.06.14.659707>
- Patching SG. Surface plasmon resonance spectroscopy for characterisation of membrane protein-ligand interactions and its potential for drug discovery. *Biochim Biophys Acta*. 2014 Jan;1838(1 Pt A):43-55. doi: 10.1016/j.bbamem.2013.04.028. Epub 2013 May 9. PMID: 23665295.
- Pereira, F. (2020). Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 85, 104525. <https://doi.org/10.1016/j.meegid.2020.104525>
- Perlinska, A. P., Stasiulewicz, A., Nguyen, M. L., Swiderska, K., Zmudzinski, M., Maksymiuk, A. W., Drag, M., & Sulkowska, J. I. (2022). Amino acid variants of SARS-CoV-2 papain-like protease have impact on drug binding. *PLOS Computational Biology*, 18(11), e1010667. <https://doi.org/10.1371/journal.pcbi.1010667>
- Peti W., Johnson M.A., Herrmann T., Neuman B.W., Buchmeier M.J., Nelson M., Joseph J., Page R., Stevens R.C., Kuhn P., Wüthrich K. Structural genomics of the severe acute respiratory syndrome coronavirus: nuclear magnetic resonance structure of the protein nsP7. *J. Virol.* 2005;79(20):12905–12913. doi: 10.1128/JVI.79.20.12905-12913.2005.
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable Molecular Dynamics with NAMD. *Journal of Computational Chemistry*, 26(16), 1781–1802. <https://doi.org/10.1002/jcc.20289>
- Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., & Lindahl, E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics (Oxford, England)*, 29(7), 845–854. <https://doi.org/10.1093/bioinformatics/btt055>
- Pyrk K., Berkhout B., van der Hoek L. The novel human coronaviruses NL63 and HKU1. *J. Virol.* 2007;81:3051–3057.
- Qing, X., Lee, X. Y., Raeymaecker, J. D., Tame, J. R., Zhang, K. Y., Maeyer, M. D., & Voet, A. R. (2014). Pharmacophore modeling: Advances, limitations, and current utility in drug discovery. *Journal of Receptor, Ligand and Channel Research*, 7, 81–92. <https://doi.org/10.2147/JRLCR.S46843>
- Rack, J. G. M., Zorzini, V., Zhu, Z., Schuller, M., Ahel, D., & Ahel, I. (2020). Viral macrodomains: a structural and evolutionary assessment of the pharmacological potential. *Open biology*, 10(11), 200237. <https://doi.org/10.1098/rsob.200237>
- Rastelli, G., Del Rio, A., Degliesposti, G., & Sgobba, M. (2010). Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *Journal of computational chemistry*, 31(4), 797–810. <https://doi.org/10.1002/jcc.21372>
- Rathore, A. P. S., Haystead, T., Das, P. K., Merits, A., Ng, M.-L., & Vasudevan, S. G. (2014). Chikungunya virus nsP3 & nsP4 interacts with HSP-90 to promote virus replication: HSP-90

- inhibitors reduce CHIKV infection and inflammation *in vivo*. *Antiviral Research*, 103, 7–16. <https://doi.org/10.1016/j.antiviral.2013.12.010>
- Ravichandran, R., & Manian, M. (2008). Ribavirin therapy for Chikungunya arthritis. *The Journal of Infection in Developing Countries*, 2(02), Article 02. <https://doi.org/10.3855/jidc.286>
- Robert X., Gouet P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* 2014;42:W320–W324.
- Robertson MJ, Tirado-Rives J, Jorgensen WL. Improved Peptide and Protein Torsional Energetics with the OPLSAA Force Field. *J Chem Theory Comput.* 2015 Jul 14;11(7):3499-509. doi: 10.1021/acs.jctc.5b00356. PMID: 26190950; PMCID: PMC4504185.
- Rogers, D. J., & Tanimoto, T. T. (1960). A Computer Program for Classifying Plants. *Science*, 132(3434), 1115–1118. <https://doi.org/10.1126/science.132.3434.1115>
- Rosas-Lemus M., Minasov G., Shuvalova L., Inniss N.L., Kiryukhina O., Wiersum G., Kim Y., Jedzejczak R., Maltseva N.I., Endres M., Jaroszewski L., Godzik A., Joachimiak A., Satchell K. The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. *bioRxiv* : Preprint Server Biol. 2020;2020 doi: 10.1101/2020.04.17.047498.04.17.047498.
- Roskoski R., Jr (2003). STI-571: an anticancer protein-tyrosine kinase inhibitor. *Biochemical and biophysical research communications*, 309(4), 709–717. <https://doi.org/10.1016/j.bbrc.2003.08.055>
- Rothan, H. A., Bahrani, H., Mohamed, Z., Teoh, T. C., Shankar, E. M., Rahman, N. A., & Yusof, R. (2015). A Combination of Doxycycline and Ribavirin Alleviated Chikungunya Infection. *PLoS ONE*, 10(5), e0126360. <https://doi.org/10.1371/journal.pone.0126360>
- Rush, T. S., 3rd, Grant, J. A., Mosyak, L., & Nicholls, A. (2005). A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of medicinal chemistry*, 48(5), 1489–1495. <https://doi.org/10.1021/jm040163o>
- Russo, L. C., Tomasin, R., Matos, I. A., Manucci, A. C., Sowa, S. T., Dale, K., Caldecott, K. W., Lehtiö, L., Schechtman, D., Meotti, F. C., Bruni-Cardoso, A., & Hoch, N. C. (2021). The SARS-CoV-2 Nsp3 macrodomain reverses PARP9/DTX3L-dependent ADP-ribosylation induced by interferon signaling. *The Journal of biological chemistry*, 297(3), 101041. <https://doi.org/10.1016/j.jbc.2021.101041>
- Rut W., Lv Z., Zmudzinski M., Patchett S., Nayak D., Snipas S.J., El Oualid F., Huang T.T., Bekes M., Drag M., Olsen S.K. Activity profiling and structures of inhibitor-bound SARS-CoV-2-PLpro protease provides a framework for anti-COVID-19 drug design. *bioRxiv* : Preprint Server Biol. 2020;2020 doi: 10.1101/2020.04.29.068890. 04.29.068890.
- Rychlewski, L., Jaroszewski, L., Li, W., & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science : A Publication of the Protein Society*, 9(2), 232–241. <https://doi.org/10.1110/ps.9.2.232>
- Saberi A., Gulyaeva A.A., Brubacher J.L., Newmark P.A., Gorbalenya A.E. A planarian nidovirus expands the limits of RNA genome size. *PLoS Pathog.* 2018;14(11) doi: 10.1371/journal.ppat.1007314.
- Sacramento C.Q., Fintelman-Rodrigues N., Temerozo J.R., Da Silva A., Dias S., da Silva C., Ferreira A.C., Mattos M., Pão C., de Freitas C.S., Soares V.C., Hoelz L., Fernandes T., Branco F., Bastos

- M.M., Boechat N., Saraiva F.B., Ferreira M.A., Jockusch S., Wang X., et al. In vitro antiviral activity of the anti-HCV drugs daclatasvir and sofosbuvir against SARS-CoV-2, the aetiological agent of COVID-19. *J. Antimicrob. Chemother.* 2021;76(7):1874–1885. doi: 10.1093/jac/dkab072.
- Saha I., Ghosh N., Pradhan A., Sharma N., Maity D., Mitra K. Whole genome analysis of more than 10 000 SARS-CoV-2 virus unveils global genetic diversity and target region of NSP6. *Briefings Bioinf.* 2021;22(2):1106–1121. doi: 10.1093/bib/bbab025.
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4), 406-425.
- Salo-Ahen, O. M. H., Alanko, I., Bhadane, R., Bonvin, A. M. J. J., Honorato, R. V., Hossain, S., Juffer, A. H., Kabedev, A., Lahtela-Kakkonen, M., Larsen, A. S., Lescrinier, E., Marimuthu, P., Mirza, M. U., Mustafa, G., Nunes-Alves, A., Patsar, T., Saadabadi, A., Singaravelu, K., & Vanmeert, M. (2021). Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development. *Processes*, 9(1), 71. <https://doi.org/10.3390/pr9010071>
- Saramago M., Bárria C., Costa V.G., Souza C.S., Viegas S.C., Domingues S., Lousa D., Soares C.M., Arraiano C.M., Matos R.G. New targets for drug design: importance of nsp14/nsp10 complex formation for the 3'-5' exoribonucleolytic activity on SARS-CoV-2. *FEBS J.* 2021;288(17):5130–5147. doi: 10.1111/febs.15815.
- Saramago M., Costa V.G., Souza C.S., Bárria C., Domingues S., Viegas S.C., Lousa D., Soares C.M., Arraiano C.M., Matos R.G. The nsp15 nuclease as a good target to combat SARS-CoV-2: mechanism of action and its inactivation with FDA-approved drugs. *Microorganisms*. 2022;10(2):342. doi: 10.3390/microorganisms10020342.
- Sasaki M., Tabata K., Kishimoto M., Itakura Y., Kobayashi H., Ariizumi T., Uemura K., Toba S., Kusakabe S., Maruyama Y., Iida S., Nakajima N., Suzuki T., Yoshida S., Nobori H., Sanaki T., Kato T., Shishido T., Hall W.W., Orba Y., et al. S-217622, a SARS-CoV-2 main protease inhibitor, decreases viral load and ameliorates COVID-19 severity in hamsters. *Sci. Transl. Med.* 2023;15(679) doi: 10.1126/scitranslmed.abq4064.
- Sayers E.W., Bolton E.E., Brister J.R., Canese K., Chan J., Comeau D.C., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B.W., Pruitt K.D., Sherry S.T. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 2022;50(D1):D20–D26. doi: 10.1093/nar/gkab1112. PMID: 34850941; PMCID: PMC8728269.
- Schrödinger, L., & DeLano, W. (2020). *PyMOL*. Retrieved from <http://www.pymol.org/pymol>
- Schubert K., Karousis E.D., Jomaa A., Scaiola A., Echeverria B., Gurzeler L.A., Leibundgut M., Thiel V., Mühlemann O., Ban N. SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nat. Struct. Mol. Biol.* 2020;27(10):959–966. doi: 10.1038/s41594-020-0511-8.
- Schuller, M., Correy, G. J., Gahbauer, S., Fearon, D., Wu, T., Díaz, R. E., Young, I. D., Carvalho Martins, L., Smith, D. H., Schulze-Gahmen, U., Owens, T. W., Deshpande, I., Merz, G. E., Thwin, A. C., Biel, J. T., Peters, J. K., Moritz, M., Herrera, N., Kratochvil, H. T., QCRG Structural Biology Consortium, ... Ahel, I. (2021). Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. *Science advances*, 7(16), eabf8711. <https://doi.org/10.1126/sciadv.abf8711>

- Seidel, T., Wieder, O., Garon, A., & Langer, T. (2020). Applications of the Pharmacophore Concept in Natural Product inspired Drug Design. *Molecular informatics*, 39(11), e2000059. <https://doi.org/10.1002/minf.202000059>
- Semper C., Watanabe N., Savchenko A. Structural characterization of nonstructural protein 1 from SARS-CoV-2. *iScience*. 2021;24(1) doi: 10.1016/j.isci.2020.101903.
- Shadrack, W. R., Ndjomou, J., Kolli, R., Mukherjee, S., Hanson, A. M., & Frick, D. N. (2013). Discovering new medicines targeting helicases: challenges and recent progress. *Journal of biomolecular screening*, 18(7), 761–781. <https://doi.org/10.1177/1087057113482586>
- Shang J., Han N., Chen Z., Peng Y., Li L., Zhou H., Ji C., Meng J., Jiang T., Wu A. Compositional diversity and evolutionary pattern of coronavirus accessory proteins. *Briefings Bioinf.* 2021;22(2):1267–1278. doi: 10.1093/bib/bbaa262.
- Shirts, M. R., Mobley, D. L., & Chodera, J. D. (2007). Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time? In D. C. Spellmeyer & R. Wheeler (Eds.), *Annual Reports in Computational Chemistry* (Vol. 3, pp. 41–59). Elsevier. [https://doi.org/10.1016/S1574-1400\(07\)03004-6](https://doi.org/10.1016/S1574-1400(07)03004-6)
- Shitrit A., Zaidman D., Kalid O., Bloch I., Doron D., Yarnizky T., Buch I., Segev I., Ben-Zeev E., Segev E., Kobilier O. Conserved interactions required for inhibition of the main protease of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) *Sci. Rep.* 2020;10(1) doi: 10.1038/s41598-020-77794-5.
- Shytaj I.L., Fares M., Gallucci L., Lucic B., Tolba M.M., Zimmermann L., Adler J.M., Xing N., Bushe J., Gruber A.D., Ambiel I., Taha Ayoub A., Cortese M., Neufeldt C.J., Stolp B., Sobhy M.H., Fathy M., Zhao M., Laketa V., Diaz R.S., Sutton R.E., Chlanda P., Boulant S., Bartenschlager R., Stanifer M.L., Fackler O.T., Trimpert J., Savarino A., Lusic M. The FDA-approved drug cobicistat synergizes with remdesivir to inhibit SARS-CoV-2 replication in vitro and decreases viral titers and disease progression in Syrian hamsters. *mBio*. 2022;13(2) doi: 10.1128/mbio.03705-21. Epub 2022 Mar 1. PMID: 35229634; PMCID: PMC8941859.
- Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.* 2011;7:539.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7, 539. <https://doi.org/10.1038/msb.2011.75>
- Silva, L. A., & Dermody, T. S. (2017). Chikungunya virus: Epidemiology, replication, disease mechanisms, and prospective intervention strategies. *The Journal of Clinical Investigation*, 127(3), 737–749. <https://doi.org/10.1172/JCI84417>
- Sinha, S., Tam, B., & Wang, S. M. (2022). Applications of Molecular Dynamics Simulation in Protein Study. *Membranes*, 12(9), 844. <https://doi.org/10.3390/membranes12090844>
- Smalley, C., Erasmus, J. H., Chesson, C. B., & Beasley, D. W. C. (2016). Status of research and development of vaccines for chikungunya. *Vaccine*, 34(26), 2976–2981. <https://doi.org/10.1016/j.vaccine.2016.03.076>

- Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195–197. [https://doi.org/10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5)
- Snijder E.J., Bredenbeek P.J., Dobbe J.C., Thiel V., Ziebuhr J., Poon L.L., Guan Y., Rozanov M., Spaan W.J., Gorbalenya A.E. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol.* 2003;331(5):991–1004. doi: 10.1016/s0022-2836(03)00865-9.
- Snijder E.J., Decroly E., Ziebuhr J. The nonstructural proteins directing coronavirus RNA synthesis and processing. *Adv. Virus Res.* 2016;96:59–126. doi: 10.1016/bs.aivir.2016.08.008.
- Sommers, J. A., Loftus, L. N., Jones, M. P., Lee, R. A., Haren, C. E., Dumm, A. J., & Brosh, R. M. (2023). Biochemical analysis of SARS-CoV-2 Nsp13 helicase implicated in COVID-19 and factors that regulate its catalytic functions. *Journal of Biological Chemistry*, 299(3), 102980. <https://doi.org/10.1016/j.jbc.2023.102980>
- Sorkun MC, Koelman JMVA, Er S. Pushing the limits of solubility prediction via quality-oriented data selection. *iScience*. 2020 Dec 17;24(1):101961. doi: 10.1016/j.isci.2020.101961. PMID: 33437941; PMCID: PMC7788089.
- Stevens LJ, Pruijssers AJ, Lee HW, Gordon CJ, Tchesnokov EP, Gribble J, George AS, Hughes TM, Lu X, Li J, Perry JK, Porter DP, Cihlar T, Sheahan TP, Baric RS, Götte M, Denison MR. Mutations in the SARS-CoV-2 RNA-dependent RNA polymerase confer resistance to remdesivir by distinct mechanisms. *Sci Transl Med.* 2022 Aug 3;14(656):eabo0718. doi: 10.1126/scitranslmed.abo0718. Epub 2022 Aug 3. PMID: 35482820; PMCID: PMC9097878.
- Stobart C.C., Sexton N.R., Munjal H., Lu X., Molland K.L., Tomar S., Mesecar A.D., Denison M.R. Chimeric exchange of coronavirus nsp5 proteases (3CLpro) identifies common and divergent regulatory determinants of protease activity. *J. Virol.* 2013;87(23):12611–12618. doi: 10.1128/JVI.02050-13.
- Strauss, E. G., Rice, C. M., & Strauss, J. H. (1983). Sequence coding for the alphavirus nonstructural proteins is interrupted by an opal termination codon. *Proceedings of the National Academy of Sciences of the United States of America*, 80(17), 5271–5275. <https://doi.org/10.1073/pnas.80.17.5271>
- Strauss, J. H., & Strauss, E. G. (1994). The alphaviruses: Gene expression, replication, and evolution. *Microbiological Reviews*, 58(3), 491–562. <https://doi.org/10.1128/mr.58.3.491-562.1994>
- Summers B.B., Beavers J.W., Klibanov O.M. Sofosbuvir, a novel nucleotide analogue inhibitor used for the treatment of hepatitis C virus. *J. Pharm. Pharmacol.* 2014;66(12):1653–1666. doi: 10.1111/jphp.12294.
- Swaim C.D., Perng Y.C., Zhao X., Canadeo L.A., Harastani H.H., Darling T.L., Boon A.C.M., Lenschow D.J., Huibregtse J.M. 6-Thioguanine blocks SARS-CoV-2 replication by inhibition of PLpro protease activities. *bioRxiv [Preprint]* 2020;2020 doi: 10.1101/2020.07.01.183020. PMID: 32637945; PMCID: PMC7337375.
- Tahir Ul Qamar M., Alqahtani S.M., Alamri M.A., Chen L.L. Structural basis of SARS-CoV-2 3CL^{pro} and anti-COVID-19 drug discovery from medicinal plants. *J. Pharmaceutic. Anal.* 2020;10(4):313–319. doi: 10.1016/j.jpha.2020.03.009.

- Tanner J.A., Zheng B.J., Zhou J., Watt R.M., Jiang J.Q., Wong K.L., Lin Y.P., Lu L.Y., He M.L., Kung H.F., Kesel A.J., Huang J.D. The adamantane-derived bananins are potent inhibitors of the helicase activities and replication of SARS coronavirus. *Chem. Biol.* 2005;12(3):303–311. doi: 10.1016/j.chembiol.2005.01.006.
- Tanner, J. A., Watt, R. M., Chai, Y. B., Lu, L. Y., Lin, M. C., Peiris, J. S., Poon, L. L., Kung, H. F., & Huang, J. D. (2003). The severe acute respiratory syndrome (SARS) coronavirus NTPase/helicase belongs to a distinct class of 5' to 3' viral helicases. *The Journal of biological chemistry*, 278(41), 39578–39582. <https://doi.org/10.1074/jbc.C300328200>
- Tazikeh-Lemeski E., Moradi S., Raoufi R., Shahlaei M., Janlou M., Zolghadri S. Targeting SARS-COV-2 non-structural protein 16: a virtual drug repurposing study. *J. Biomol. Struct. Dyn.* 2020 doi: 10.1080/07391102.2020.1779133.
- Thiel V., Ivanov K.A., Putics Á., Hertzog T., Schelle B., Bayer S., Weißbrich B., Snijder E.J., Rabenau H., Doerr H.W., Gorbalenya A.E., Ziebuhr J. Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* 2003;84(Pt 9):2305–2315. doi: 10.1099/vir.0.19424-0.
- Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22), 4673–4680. <https://doi.org/10.1093/nar/22.22.4673>
- Thompson, J. D., Linard, B., Lecompte, O., & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one*, 6(3), e18093.
- Tok T.T., Tatar G. Structures and functions of coronavirus proteins: molecular modeling of viral nucleoprotein. *Int. J. Virol. Infect. Dis.* 2017;2 001–006.
- Troshin P.V., Procter J.B., Barton G.J. Java bioinformatics analysis web services for multiple sequence alignment–JABAWS:MSA. *Bioinformatics*. 2011;27:2001–2002. doi: 10.1093/bioinformatics/btr304.
- Troshin P.V., Procter J.B., Sherstnev A., Barton D.L., Madeira F., Barton G.J. JABAWS 2.2 distributed web services for bioinformatics: protein disorder, conservation and RNA secondary structure. *Bioinformatics*. 2018;34:1939–1940. doi: 10.1093/bioinformatics/bty045.
- Turner, P.J. (2005). XMGRACE, Version 5.1. 19. Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR, 2.5 (2005): 19.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352. <https://doi.org/10.1037/0033-295X.84.4.327>
- Umate, P., Tuteja, N., & Tuteja, R. (2011). Genome-wide comprehensive analysis of human helicases. *Communicative & integrative biology*, 4(1), 118–137. <https://doi.org/10.4161/cib.4.1.13844>
- Unal M.A., Bitirim C.V., Summak G.Y., Bereketoglu S., Cevher Zeytin I., Besbinar O., Gurcan C., Aydos D., Goksoy E., Kocakaya E., Eran Z., Murat M., Demir N., Aksoy Ozer Z.B., Somers J., Demir E., Nazir H., Ozkan S.A., Ozkul A., Azap A., et al. Ribavirin shows antiviral activity against SARS-CoV-2 and downregulates the activity of TMPRSS2 and the expression of ACE2 in vitro. *Can. J. Physiol. Pharmacol.* 2021;99(5):449–460. doi: 10.1139/cjpp-2020-0734.

- Unoh Y., Uehara S., Nakahara K., Nobori H., Yamatsu Y., Yamamoto S., Maruyama Y., Taoda Y., Kasamatsu K., Suto T., Kouki K., Nakahashi A., Kawashima S., Sanaki T., Toba S., Uemura K., Mizutare T., Ando S., Sasaki M., Orba Y., et al. Discovery of S-217622, a noncovalent oral SARS-CoV-2 3CL protease inhibitor clinical candidate for treating COVID-19. *J. Med. Chem.* 2022;65(9):6499–6512. doi: 10.1021/acs.jmedchem.2c00117.
- Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A., & Moreno, E. (2021). gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *Journal of Chemical Theory and Computation*, 17(10), 6281–6291. <https://doi.org/10.1021/acs.jctc.1c00645>
- Van Boheemen S., De Graaf M., Lauber C., Bestebroer T., Raj V.S., Zaki A., Osterhaus A., Haagmans B., Gorbelenya A., Snijder E., et al. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *mBio*. 2012;3doi: 10.1128/mBio.00473-1.
- Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. C. (2005). GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16), 1701–1718. <https://doi.org/10.1002/jcc.20291>
- van Duijl-Richter, M. K. S., Blijleven, J. S., van Oijen, A. M., & Smit, J. M. (2015). Chikungunya virus fusion properties elucidated by single-particle and bulk approaches. *Journal of General Virology*, 96(8), 2122–2132. <https://doi.org/10.1099/vir.0.000144>
- Van Gunsteren, W. F., & Berendsen, H. J. C. (1988). A Leap-frog Algorithm for Stochastic Dynamics. *Molecular Simulation*, 1(3), 173–185. <https://doi.org/10.1080/08927028808080941>
- Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem.* 2010 Mar;31(4):671-90. doi: 10.1002/jcc.21367. PMID: 19575467; PMCID: PMC2888302.
- Verba K., Gupta M., Azumaya C., Moritz M., Pourmal S., Diallo A., Merz G., Jang G., Bouhaddou M., Fossati A., Brilot A., Diwanji D., Hernandez E., Herrera N., Kratochvil H., Lam V., Li F., Li Y., Nguyen H., Nowotny C., et al. CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. *Res. Square*. 2021;3 doi: 10.21203/rs.3.rs-515215/v1. rs-515215.
- von Grotthuss M., Wyrwicz L.S., Rychlewski L. mRNA cap-1 methyltransferase in the SARS genome. *Cell*. 2003;113(6):701–702. doi: 10.1016/s0092-8674(03)00424-0.
- Voss, J. E., Vaney, M.-C., Duquerroy, S., Vonnheim, C., Girard-Blanc, C., Crublet, E., Thompson, A., Bricogne, G., & Rey, F. A. (2010). Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. *Nature*, 468(7324), 709–712. <https://doi.org/10.1038/nature09555>
- Wahid, B., Ali, A., Rafique, S., & Idrees, M. (2017). Global expansion of chikungunya virus: Mapping the 64-year history. *International Journal of Infectious Diseases*, 58, 69–76. <https://doi.org/10.1016/j.ijid.2017.03.006>
- Wan Y., Shang J., Graham R., Baric R.S., Li F. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J. Virol.* 2020;94(7):e00127–20. doi: 10.1128/JVI.00127-20.

- Wang N., Shi X., Jiang L., Zhang S., Wang D., Tong P., Guo D., Fu L., Cui Y., Liu X., Arledge K.C., Chen Y.H., Zhang L., Wang X. Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4. *Cell Res.* 2013;23(8):986–993. doi: 10.1038/cr.2013.92.
- Wang, J., Deng, Y., & Roux, B. (2006). Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophysical journal*, 91(8), 2798–2814. <https://doi.org/10.1529/biophysj.106.084301>
- Wang, X., Han, W., Yan, X., Zhang, J., Yang, M., & Jiang, P. (2020). Pharmacophore features for machine learning in pharmaceutical virtual screening. *Molecular diversity*, 24(2), 407–412. <https://doi.org/10.1007/s11030-019-09961-4>
- Wang, Y., & Li, K.-B. (2004). An adaptive and iterative algorithm for refining multiple sequence alignment. *Computational Biology and Chemistry*, 28(2), 141–148. <https://doi.org/10.1016/j.compbiolchem.2004.02.001>
- Weaver, S. C., & Lecuit, M. (2015). Chikungunya Virus and the Global Spread of a Mosquito-Borne Disease. *New England Journal of Medicine*, 372(13), 1231–1239. <https://doi.org/10.1056/NEJMra1406035>
- Weaver, S. C., Osorio, J. E., Livengood, J. A., Chen, R., & Stinchcomb, D. T. (2012). Chikungunya virus and prospects for a vaccine. *Expert Review of Vaccines*, 11(9), 1087–1101. <https://doi.org/10.1586/erv.12.84>
- Weber, S., Ramirez, C. M., Weiser, B., Burger, H., & Doerfler, W. (2021). SARS-CoV-2 worldwide replication drives rapid rise and selection of mutations across the viral genome: A time-course study – potential challenge for vaccines and therapies. *EMBO Molecular Medicine*, 13(6), e14062. <https://doi.org/10.15252/emmm.202114062>
- Wermuth, C., Ganellin, C., Lindberg, P. & Mitscher, L. (1998). Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure and Applied Chemistry*, 70(5), 1129–1143. <https://doi.org/10.1351/pac199870051129>
- White M.A., Lin W., Cheng X. Discovery of COVID-19 inhibitors targeting the SARS-CoV2 Nsp13 helicase. *bioRxiv* : Preprint Server Biol. 2020;2020 doi: 10.1101/2020.08.09.243246.08.09.243246.
- WHO Solidarity Trial Consortium. Pan H., Peto R., Henao-Restrepo A.M., Preziosi M.P., Sathiyamoorthy V., Abdool Karim Q., Alejandria M.M., Hernández García C., Kieny M.P., Malekzadeh R., Murthy S., Reddy K.S., Roses Periago M., Abi Hanna P., Ader F., Al-Bader A.M., Alhasawi A., Allum E., Alotaibi A., et al. Repurposed antiviral drugs for covid-19 - interim WHO solidarity trial results. *N. Engl. J. Med.* 2021;384(6):497–511. doi: 10.1056/NEJMoa2023184.
- Wilbur, W. J., & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences*, 80(3), 726–730.
- Williamson B.N., Feldmann F., Schwarz B., Meade-White K., Porter D.P., Schulz J., van Doremalen N., Leighton I., Yinda C.K., Pérez-Pérez L., Okumura A., Lovaglio J., Hanley P.W., Saturday G., Bosio C.M., Anzick S., Barbican K., Cihlar T., Martens C., Scott D.P., et al. Clinical benefit of remdesivir in rhesus macaques infected with SARS-CoV-2. *Nature*. 2020;585(7824):273–276. doi: 10.1038/s41586-020-2423-5.
- Wolber, G., & Sippl, W. (2015). Pharmacophore Identification and Pseudo-Receptor Modeling. In *The Practice of Medicinal Chemistry* (pp. 489–510). Elsevier. <https://doi.org/10.1016/B978-0-12-417205-0.00021-3>

- Wong C.K.H., Au I.C.H., Lau K.T.K., Lau E.H.Y., Cowling B.J., Leung G.M. Real-world effectiveness of early molnupiravir or nirmatrelvir-ritonavir in hospitalised patients with COVID-19 without supplemental oxygen requirement on admission during Hong Kong's omicron BA.2 wave: a retrospective cohort study. *Lancet Infect. Dis.* 2022;22(12):1681–1693. doi: 10.1016/S1473-3099(22)00507-2.
- Woo P.C.Y., Lau S.K.P., Lam C.S.F., Lau C.C.Y., Tsang A.K.L., Lau J.H.N., Bai R., Teng J.L.L., Tsang C.C.C., Wang M., et al. Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J. Virol.* 2012;86:3995–4008.
- Wu A., Peng Y., Huang B., Ding X., Wang X., Niu P., Meng J., Zhu Z., Zhang Z., Wang J., Sheng J., Quan L., Xia Z., Tan W., Cheng G., Jiang T. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe.* 2020;27(3):325–328. doi: 10.1016/j.chom.2020.02.001.
- Wu C., Liu Y., Yang Y., Zhang P., Zhong W., Wang Y., Wang Q., Xu Y., Li M., Li X., Zheng M., Chen L., Li H. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sin. B.* 2020;10(5):766–788. doi: 10.1016/j.apsb.2020.02.008.
- Wu S, Manber U: Fast Text Searching Allowing Errors. *Communications of the ACM* 1992, 35: 83–91
- Xu J., Zhao S., Teng T., Abdalla A.E., Zhu W., Xie L., Wang Y., Guo X. Systematic comparison of two animal-to-human transmitted human coronaviruses: SARS-CoV-2 and SARS-CoV. *Viruses.* 2020;12(2):244. doi: 10.3390/v12020244.
- Xu X., Chen P., Wang J., Feng J., Zhou H., Li X., Zhong W., Hao P. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci. China Life Sci.* 2020;63(3):457–460. doi: 10.1007/s11427-020-1637-5.
- Xu X., Liu Y., Weiss S., Arnold E., Sarafianos S.G., Ding J. Molecular model of SARS coronavirus polymerase: implications for biochemical functions and drug design. *Nucleic Acids Res.* 2003;31(24):7117–7130. doi: 10.1093/nar/gkg916.
- Yazdani S., De Maio N., Ding Y., Shahani V., Goldman N., Schapira M. Genetic variability of the SARS-CoV-2 pocketome. *J. Proteome Res.* 2021;20(8):4212–4215. doi: 10.1021/acs.jproteome.1c00206.
- Yazdi, A. K., Pakarian, P., Perveen, S., Hajian, T., Santhakumar, V., Bolotokova, A., Li, F., & Vedadi, M. (2022). Kinetic Characterization of SARS-CoV-2 nsp13 ATPase Activity and Discovery of Small-Molecule Inhibitors. *ACS infectious diseases*, 8(8), 1533–1542. <https://doi.org/10.1021/acsinfecdis.2c00165>
- Yeager C.L., Ashmun R.A., Williams R.K., Cardellicchio C.B., Shapiro L.H., Look A.T., Holmes K.V. Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature.* 1992;357(6377):420–422. doi: 10.1038/357420a0.
- Yin W., Mao C., Luan X., Shen D.D., Shen Q., Su H., Wang X., Zhou F., Zhao W., Gao M., Chang S., Xie Y.C., Tian G., Jiang H.W., Tao S.C., Shen J., Jiang Y., Jiang H., Xu Y., Zhang S., Zhang Y., Xu H.E. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by

- remdesivir. *Science*. 2020;368(6498):1499–1504. doi: 10.1126/science.abc1560. Epub 2020 May 1. PMID: 32358203; PMCID: PMC7199908.
- Yoneyama M., Fujita T. Recognition of viral nucleic acids in innate immunity. *Rev. Med. Virol.* 2010;20(1):4–22. doi: 10.1002/rmv.633.
- Zanotto P.M., Gibbs M.J., Gould E.A., Holmes E.C. A reevaluation of the higher taxonomy of viruses based on RNA polymerases. *J. Virol.* 1996;70(9):6083–6096. doi: 10.1128/JVI.70.9.6083-6096.1996.
- Zeng J., Weissmann F., Bertolin A.P., Posse V., Canal B., Ulferts R., Wu M., Harvey R., Hussain S., Milligan J.C., Roustan C., Borg A., McCoy L., Drury L.S., Kjaer S., McCauley J., Howell M., Beale R., Diffley J. Identifying SARS-CoV-2 antiviral compounds by screening for small molecule inhibitors of nsp13 helicase. *Biochem. J.* 2021;478(13):2405–2423. doi: 10.1042/BCJ20210201.
- Zhang H., Penninger J.M., Li Y., Zhong N., Slutsky A.S. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Med.* 2020;46(4):586–590. doi: 10.1007/s00134-020-05985-9.
- Zhang L., Liu J., Cao R., Xu M., Wu Y., Shang W., Wang X., Zhang H., Jiang X., Sun Y., Hu H., Li Y., Zou G., Zhang M., Zhao L., Li W., Guo X., Zhuang X., Yang X.L., Shi Z.L., Deng F., Hu Z., Xiao G., Wang M., Zhong W. Comparative antiviral efficacy of viral protease inhibitors against the novel SARS-CoV-2 in vitro. *Virol. Sin.* 2020;35(6):776–784. doi: 10.1007/s12250-020-00288-1. Epub 2020 Sep 10. PMID: 32910347; PMCID: PMC7481761.
- Zhang, K., Law, M. C. Y., Nguyen, T. M., Tan, Y. B., Wirawan, M., Law, Y.-S., Jeong, L. S., & Luo, D. (2022). Molecular basis of specific viral RNA recognition and 5'-end capping by the Chikungunya virus nsP1. *Cell Reports*, 40(4), 111133. <https://doi.org/10.1016/j.celrep.2022.111133>
- Zhang, R., Hryc, C. F., Cong, Y., Liu, X., Jakana, J., Gorchakov, R., Baker, M. L., Weaver, S. C., & Chiu, W. (2011). 4.4 Å cryo-EM structure of an enveloped alphavirus Venezuelan equine encephalitis virus. *The EMBO Journal*, 30(18), 3854–3863. <https://doi.org/10.1038/emboj.2011.261>
- Zhang, Y., Kandwal, S., Fayne, D., & Stevenson, N. J. (2024). MERS-CoV-nsp5 expression in human epithelial BEAS 2b cells attenuates type I interferon production by inhibiting IRF3 nuclear translocation. *Cellular and molecular life sciences : CMLS*, 81(1), 433. <https://doi.org/10.1007/s00018-024-05458-y>
- Zhang, Y., Zhang, Q., Zhou, J., & Zou, Q. (2022). A survey on the algorithm and development of multiple sequence alignment. *Briefings in Bioinformatics*, 23(3), bbac069. <https://doi.org/10.1093/bib/bbac069>
- Zhou P., Yang X.L., Wang X.G., Hu B., Zhang L., Zhang W., Si H.R., Zhu Y., Li B., Huang C.L., Chen H.D., Chen J., Luo Y., Guo H., Jiang R.D., Liu M.Q., Chen Y., Shen X.R., Wang X., Zheng X.S., et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270–273. doi: 10.1038/s41586-020-2012-7.
- Zhou, Z. J., Qiu, Y., Pu, Y., Huang, X., & Ge, X. Y. (2020). BioAider: An efficient tool for viral genome analysis and its application in tracing SARS-CoV-2 transmission. *Sustainable cities and society*, 63, 102466. <https://doi.org/10.1016/j.scs.2020.102466>

Zumla A., Chan J.F., Azhar E.I., Hui D.S., Yuen K.Y. Coronaviruses - drug discovery and therapeutic options. *Nature Reviews. Drug Dis.* 2016;15(5):327–347. doi: 10.1038/nrd.2015.37.

Zwanzig, R. W. (1954). High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics*, 22(8), 1420–1426. <https://doi.org/10.1063/1.1740409>